Modelling the galaxy population

Christopher D. Rimes

Doctor of Philosophy

The University of Edinburgh

2003

This thesis is my own composition except where indicated in the text.

Christopher D. Rimes

21 August, 2003

Acknowledgements

I am extremely grateful to my supervisors, Eelco van Kampen and John Peacock, for all their time and assistance over the last 4 years and for reading and commenting on the many earlier drafts of this thesis. Thanks to Rob Smith, for providing the code for Peacock & Smith (2000) and for guidance with the implementation; also to Louisa Nolan for help with stellar population models. On a personal note, I would like to thank my family for their support over the years and especially during the last few months. Finally, thanks go to the occupants, both past and present, of the *Remote Observing Room*, for all the distractions.

Chris Rimes

21 August, 2003

Abstract

I describe a combined N-body/phenomenological model for the formation and evolution of galaxies in a hierarchical universe. The aim of this work is twofold: to further our understanding of the galaxy formation process and to investigate how selection effects bias our view of the galaxy population.

The cold dark matter (CDM) model for structure formation has been remarkably successful at explaining the large-scale properties of the universe and now forms part of the standard picture of galaxy formation. That the matter content of the universe is dominated by a non-luminous (dark), non-baryonic form of matter is commonly accepted; its existence is neccessary to explain observed mass-to-light ratios of structures on all scales. The evolution of dark matter on large scales has been studied extensively using N-body techniques and the results are in good agreement with the observed large-scale structure of the universe. However, our view of the universe is biased by the fact that we can only directly observe luminous (baryonic) matter, which is concentrated in stars and galaxies. To gain a complete understanding of the universe it is therefore vital to have a model for the relationship between the properties of the observed galaxy population and the underlying dark matter distribution.

Semi-analytical and related models are an attempt to bring together all of the ingredients neccessary for galaxy formation (dark matter evolution, gas dynamics, star formation and evolution, etc.) in a single model with the power to predict the properties of statistically significant numbers of galaxies. The original models used an analytical approach (hence the name) to modelling dark matter evolution, which contained no information on the distribution of galaxies. A recent development has been the coupling of similar models with N-body simulations, allowing galaxy formation to be understood

in its full cosmological context.

Another way in which our view of the universe is biased is by the limits placed on observational samples of galaxies. Galaxies are often selected for inclusion in a sample on the basis of their apparent magnitude and such samples are often termed 'magnitude-limited' samples. However, all observations have an isophotal limit, imposed by the surface brightness of the night sky, below which galaxies cannot be seen regardless of their total flux. Such low surface-brightness galaxies make an as yet unquantified contribution to the number density of galaxies in the universe, so it is important to take these into account when interpreting observations and when comparing the observed properties of the galaxy population to the predictions of galaxy formation models.

In this thesis, I present a model for galaxy formation which makes use of a customised N-body simulation with a built-in recipe for the formation of galaxy haloes. The specific aim of this work is to model the formation and evolution of disc galaxies with a view to reproducing the observed distribution of galaxies over luminosity and surface brightness. I then use these results to investigate the role of selection effects by constructing mock surveys, an application that will find much use in interpreting the results of recent and future galaxy surveys.

Contents

Chapter 1

Background

1.1 Expansion of the universe

The dynamics of the expanding universe are governed by Friedmann's equations:

$$\dot{R}^2 - \frac{8\pi G\rho}{3}R^2 = -kc^2 \tag{1.1}$$

$$\ddot{R} = -\frac{4\pi GR}{3}\left(\rho + \frac{3p}{c^2}\right), \tag{1.2}$$

which describe the evolution of the scale factor, R, of the universe. Here, ρ is the total density of the universe, p is the pressure and $k = -1$, 0 and $+1$ for universes with open (negative curvature), flat and closed (positive curvature) geometries respectively. These equations are arrived at by inserting the metric for a homogeneous, isotropic universe — the Robertson-Walker metric — into Einstein's field equations. The Robertson-Walker metric can be written as

$$c^2\mathrm{d}\tau^2 = c^2\mathrm{d}t^2 - R^2(t)[\mathrm{d}r^2 + S_k^2(r)\mathrm{d}\psi^2], \tag{1.3}$$

with

$$S_k(r) = \begin{cases} \sin r & k = 1 \\ r & k = 0 \\ \sinh r & k = -1 \end{cases} . \tag{1.4}$$

At this point, it is convenient to define a dimensionless scale factor, $a = R/R_0$, where R_0 is the present-day value of R, and a dimensionless expansion rate

$$H \equiv \frac{\dot{a}}{a}, \tag{1.5}$$

which is equivalent to the constant of proportionality in Hubble's law for the recessional velocities of galaxies ($v = H_0 d$) at small separations. Using these two definitions, equation 1.1 can be re-written as

$$H^2 - \frac{8\pi G\rho}{3} = -\frac{kc^2}{a^2 R_0^2}. \tag{1.6}$$

Clearly, there is a critical value of the density,

$$\rho_c \equiv \frac{3H^2}{8\pi G}, \tag{1.7}$$

for which $k = 0$ and the universe has a flat geometry.

The known constituents of the universe can be conveniently divided into three types, according to their equations of state:

1. non-relativistic matter ($p_M = 0$, $\rho_M \propto a^{-3}$),
2. relativistic matter and radiation ($p_R = \frac{1}{3}\rho_R c^2$, $\rho_R \propto a^{-4}$),
3. vacuum energy, or the cosmological constant, ($p_\Lambda = -\rho_\Lambda c^2$, $\rho_\Lambda \sim$ constant).[1]

Here, the label non-relativistic matter includes both dark matter and baryons. It is useful to write the densities of these three components as fractions of the critical density at a given epoch:

$$\Omega_M(a) \equiv \frac{\rho_M}{\rho_c} = \frac{8\pi G\rho_M}{3H^2}, \tag{1.8}$$

$$\Omega_R(a) \equiv \frac{\rho_R}{\rho_c} = \frac{8\pi G\rho_R}{3H^2}, \tag{1.9}$$

$$\Omega_\Lambda(a) \equiv \frac{\rho_\Lambda}{\rho_c} = \frac{8\pi G\rho_\Lambda}{3H^2}. \tag{1.10}$$

From now on, unless the dependence on a is given explicitly, I reserve the symbols Ω_M, Ω_R and Ω_Λ to denote the *present-day* densities of the three components. Inserting equations (1.8) – (1.10) and their a-dependencies into Friedmann's equation in the form given in equation (1.6), an expression is obtained for the evolution of the Hubble parameter with scale factor:

$$H^2(a) = H_0^2 \left[(1-\Omega)a^{-2} + \Omega_\Lambda + \Omega_M a^{-3} + \Omega_R a^{-4}\right], \tag{1.11}$$

[1] The possibility of a fourth type, *quintessence*, which has an equation of state $p_Q = w\rho_Q c^2$ ($-1 \leqslant w \leqslant 0$), is discussed briefly later.

where $\Omega = \Omega_{\rm M} + \Omega_{\rm R} + \Omega_\Lambda$ is the total density of the universe. For the simple cases of flat ($\Omega = 1$), matter-dominated and radiation-dominated universes, this can easily be solved to give

$$a(t) = \begin{cases} (t/t_0)^{2/3} & \text{matter-dominated} \\ (t/t_0)^{1/2} & \text{radiation-dominated} \end{cases}, \tag{1.12}$$

where t_0 is the present time. For more general cosmologies, including those with vacuum energy, equation (1.11) must be integrated numerically.

It is useful to write equation (1.11) in terms of observable quantities. If a galaxy is observed with a redshift $z(\equiv \Delta\lambda/\lambda$, where λ is the rest wavelength of a feature in the galaxy's spectrum and $\Delta\lambda$ is the observed shift in wavelength) then it is easy to show that $1 + z = 1/a_{\rm obs}$, where $a_{\rm obs}$ is the scale factor of the universe when the observed light was emitted by the galaxy (again normalised to unity at the present day). So the evolution of the universe can be described in terms of observable variables by the expression

$$H^2(z) = H_0^2 \left[(1-\Omega)(1+z)^2 + \Omega_\Lambda + \Omega_{\rm M}(1+z)^3 + \Omega_{\rm R}(1+z)^4\right]. \tag{1.13}$$

This is plotted in Fig. 1.1(a) for several commonly adopted cosmological models: a matter-dominated, flat universe (SCDM: $\Omega_{\rm M} = 1$, $\Omega_\Lambda = 0$); a vacuum-dominated, flat universe (ΛCDM: $\Omega_\Lambda = 0.7$, $\Omega_{\rm M} = 0.3$) and an open universe (OCDM: $\Omega_{\rm M} = 0.3$, $\Omega_\Lambda = 0$). The first of these was, until recently, the standard model (hence SCDM), but the ΛCDM model is favoured by current observations; the last model is almost certainly ruled out and is included for illustrative purposes only.

A dimensionless deceleration parameter,

$$q \equiv -\frac{\ddot{a}a}{\dot{a}^2}, \tag{1.14}$$

can also be defined, enabling us to re-write equation (1.2) as

$$q = \frac{4\pi G}{3H^2}\left(\rho + \frac{3p}{c^2}\right). \tag{1.15}$$

The conditions $q > 0$ and $q < 0$ correspond to cases in which the expansion of the universe is slowing down and speeding up respectively. Using equations (1.8) – (1.10) and the equations of state of the three components of the density, the evolution of the

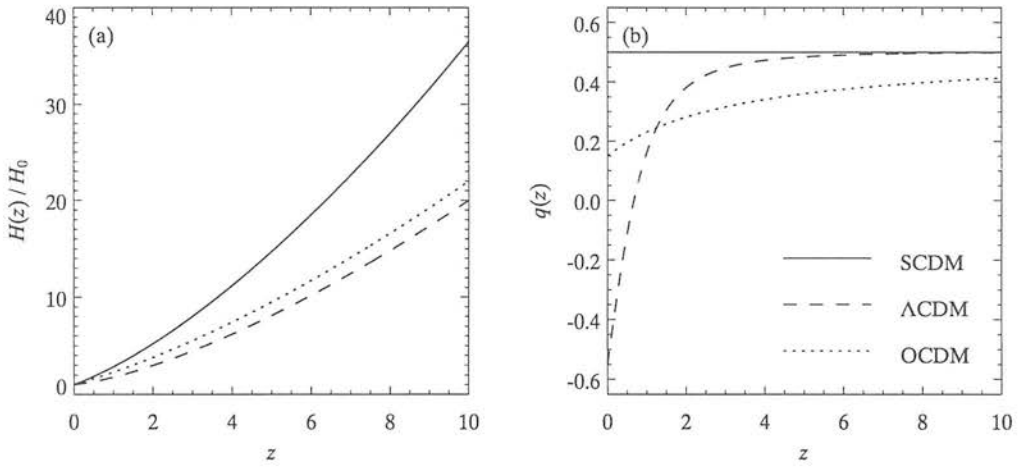

Figure 1.1: Evolution with redshift of (a) the Hubble parameter and (b) the deceleration parameter for 3 model universes: a matter-dominated, flat universe (SCDM, solid line); a vacuum-dominated, flat universe (ΛCDM, dashed line) and an open universe (OCDM, dotted line). Only the vacuum-dominated universe is consistent with an accelerating expansion ($q < 0$) at the present day.

deceleration parameter can be written as

$$q(a) = \frac{\Omega_{\rm M}(a)}{2} + \Omega_{\rm R}(a) - \Omega_\Lambda(a). \tag{1.16}$$

This is plotted in Fig. 1.1(b) for the three cosmological models described above. Ignoring the contribution from relativistic matter, which is negligible at the present day ($\Omega_{\rm R} \simeq 4.2 \times 10^{-5} h_0^{-2}$ [2], estimated from the temperature of the cosmic microwave background and assuming three species of massless neutrino with a density 0.68 times that of radiation; Peacock 1999), the condition for the universe to be accelerating at the present day is $\Omega_\Lambda > \Omega_{\rm M}/2$. An accelerating expansion is strongly suggested by observations of distant supernovae (Perlmutter et al. 1997, 1999), which, it is believed, can be used as 'standard candles' to measure the distances to external galaxies.

In Fig. 1.2 I illustrate the evolution with redshift of the densities of non-relativistic matter, relativistic matter and vacuum energy, for the three cosmological models described above. Radiation and other relativistic particles dominate the energy density of

[2]The notation h_0 is commonly used to denote the present-day value of the Hubble parameter, in units of 100 km s^{-1} Mpc^{-1}.

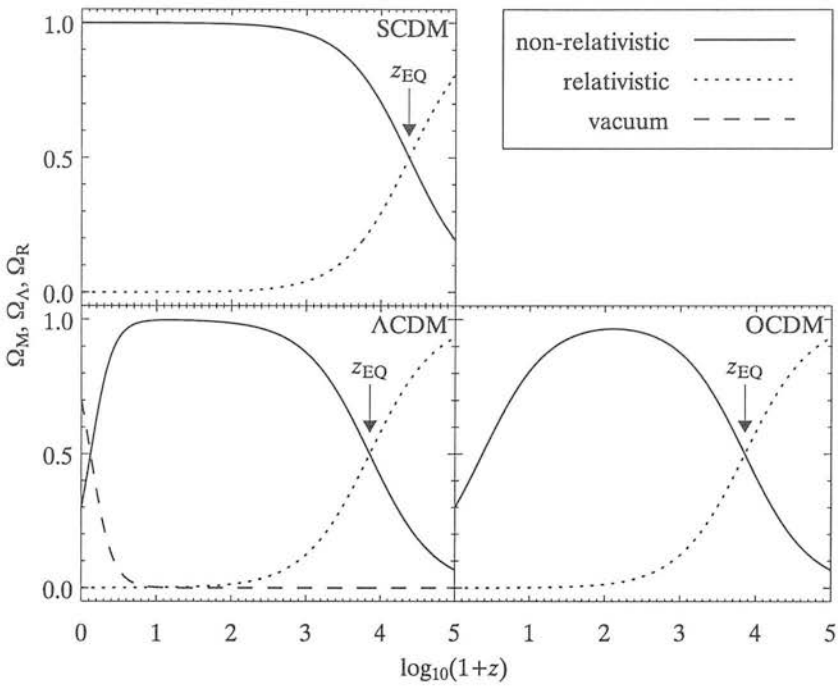

Figure 1.2: Densities of the major components of the universe as a function of redshift, for the same three model universes as Fig. 1.1. Solid lines denote non-relativistic matter, dotted lines relativistic matter and radiation and the dashed line vacuum energy. Matter-radiation equality occurs at $z_{EQ} \simeq 10^4$ in models with $\Omega_M = 0.3$, earlier in the model with $\Omega_M = 1$.

the universe at high redshifts but because, unlike non-relativistic particles, their energy decreases with the expansion of the universe, the low-redshift universe is dominated by non-relativistic matter. The exception is the universe with a cosmological constant, where vacuum energy dominates at low redshifts. The transition from a radiation-dominated universe to one dominated by matter occurs at $z_{EQ} \simeq 10^4$, the so-called 'epoch of matter-radiation equality'.

Note on Quintessence

There has recently been considerable interest in cosmological models which incorporate a fourth energy component: quintessence is defined by an equation of state $p_Q = w\rho_Q c^2$, with $-1 \leqslant w \leqslant 0$, and a density which varies as $\rho_Q \propto a^{-3(1+w)}$. The vacuum energy is a special case of quintessence with $w = -1$. Quintessence models also lead to an

accelerating expansion at the present day and, unlike the vacuum energy, quintessence is allowed to be clustered on large scales, which has consequences for structure formation. Present observations are consistent with $w = -1$ (i.e. the standard cosmological constant), but with large error bars (e.g. Spergel et al. 2003)

1.2 Hierarchical structure formation

1.2.1 The CDM universe

In the standard picture, galaxies formed through gravitational amplification of density fluctuations in the early universe. It is now commonly accepted that the matter content of the universe is dominated by a non-luminous (dark), non-baryonic form of matter. The existence of dark matter on all scales is required to explain the observed mass-to-light ratios of galaxies and larger structures. HI measurements of the rotation curves of galaxies, which extend far beyond the optical disc, imply typical mass-to-light ratios for galaxies of $M/L \simeq 30h_0$ in solar units (Peacock 1999), with much higher values in the outer regions of discs, suggesting that the visible (stellar) mass is a small fraction of the total. Estimates of the masses of galaxy clusters, from dynamical arguments (e.g. The & White 1986) and gravitational lensing (Miralda-Escude & Babul 1995), yield values of $M/L \simeq 300h_0$ (Brainerd et al. 1999) — an order of magnitude more than can be accounted for by the visible galaxies (even assuming that these contain dark matter as discussed above) and hot intracluster gas. Big bang nucleosynthesis places a constraint on the density of baryons in the universe: $\Omega_{\rm baryon} \simeq 0.01 - 0.03h_0^{-2}$ (Walker et al. 1991; Tytler, Fan & Burles 1996; Burles & Tytler 1998). If the mass-to-light ratios of galaxy clusters, the largest virialised structures in the universe, are typical of the universe as a whole then the measured luminosity density ($j_{\rm B} \simeq 1$–2×10^8 L$_\odot$ Mpc^{-3}; Cross & Driver 2002) in the local universe implies a matter density of $\Omega_{\rm M} \simeq 0.2$ and we can see immediately that a significant fraction of the matter in the universe must be of a non-baryonic form. The properties of this non-baryonic dark matter can have important consequences for the process of structure formation.

Cold dark matter (CDM) is the generic name given to particles that are non-relativistic when they decouple from the rest of the matter and radiation early in the

universe. Candidates include supersymmetric (SUSY) particles, such as the gravitino and the photino — the existences of which are predicted by some extensions to the standard model of particle physics — and other massive ($m_{\rm CDM} \gtrsim 1$ GeV) subatomic particles, collectively known as *weakly-interacting massive particles* (WIMPs). The exact nature of the particles is unimportant from the point of view of the general properties of the model. The mass within the horizon at decoupling is small ($\ll M_\odot$) and, since the particles are non-relativistic at this point, damping by free-streaming is unimportant. In CDM cosmologies, the amplitudes of perturbations increase with decreasing scale so that small-scale perturbations collapse first and, after recombination, the first structures to form are on galaxy and sub-galaxy scales. These then merge into larger structures in a process of hierarchical, or 'bottom-up', structure formation. This type of model predicts that the first generation of stars forms early in the universe, before galaxies and galaxy clusters. An alternative theory — the 'top-down' model, in which the matter content of the universe is dominated by a light ($\lesssim 10$ ev) species of subatomic particle (candidates are known collectively as hot dark matter) that is highly relativistic when it decouples — is almost certainly ruled out by observations and is no longer considered as a viable model for structure formation. Its most important feature is that stars and galaxies form late in the universe, by fragmentation of larger structures.

For a long time the CDM model has been the most successful at matching observations of the universe. In particular it explains the early re-ionisation of the intergalactic medium (IGM) by the first generation of stars and the chemical enrichment of intergalactic gas at high redshifts. Neither of these observations can be explained in the context of a top-down theory of galaxy formation. Hierarchical structure formation also has the advantage that it involves only gravity, the physics of which is well understood, and is therefore easy to model, either analytically or using N-body simulations.

However, the CDM model also suffers from several problems: it predicts an excess of dwarf galaxies over the numbers observed and — a related issue — it cannot easily explain the observed angular momenta of disc galaxies. One solution to the problem of an excess of structure on small scales is warm dark matter (WDM). This consists of hypothetical particles with masses of $m_{\rm WDM} \sim 1$ keV — more massive than typical

HDM candidates, but less massive than WIMPs. They are still ultra-relativistic when they decouple from the rest of the universe at $z \sim 10^{13}$, so that free-streaming is important and perturbations on scales $\lesssim 10^{10} - 10^{12}$ $M_{\odot}$ are severely damped, while larger scale structure behaves in much the same way as in the CDM model. This possibility seems to be ruled out by the detection of re-ionisation at $z \simeq 20$ (Spergel et al. 2003), which requires small structures to have collapsed early in the universe. An alternative to WDM is to assume a mixture of hot and cold dark matter — often called mixed dark matter (MDM) — with the relative amounts adjusted to erase structure on the right scales. This model could be realised if the dominant form of dark matter was cold but, in addition, neutrinos were found to have a finite rest mass.

1.2.2 Linear growth of perturbations

INITIAL CONDITIONS

The equations in section 1.1 describe the evolution of a homogeneous universe. While observations confirm that the universe is indeed homogeneous on the largest scales, structure is observed on scales ranging from stars and galaxies ($\lesssim$ 100 kpc) to the networks of filaments seen in large galaxy surveys ($\sim$ 100 Mpc). For these structures to form, small perturbations to the density field in the early universe are required. Over-dense regions are enhanced by gravitational infall and become the potential wells in which the baryons settle to form the galaxies we see.

The ultimate source of the perturbations present in the early universe is still a matter for debate. A common theory is one in which quantum fluctuations in the energy density of the early universe were stretched by a period of inflation, during which the scale factor of the universe grew exponentially. As well as providing an explanation for the origins of structure in the universe, this theory also explains several other important observations: (i) the large-scale isotropy of the microwave background, since regions that are in causal contact before inflation begins are stretched to scales much larger than the event horizon by the expansion, and (ii) the observed flatness of the universe, since a period of rapid expansion, in which the scale factor of the universe increases to many times its original value, will flatten out any curvature that

existed before the expansion began. The simplest inflation models predict a Gaussian distribution of perturbations (that is, a distribution whose properties are completely determined by its power spectrum) and this expectation is borne out by the observed fluctuations in the temperature of the cosmic microwave background (Komatsu et al. 2003).

Equations of linear growth

The motion of the cosmic fluid is governed by three equations:

$$\frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} = -\frac{\nabla_x p}{\rho} - \nabla_x \Phi, \tag{1.17}$$

$$\frac{\mathrm{D}\rho}{\mathrm{D}t} = -\rho \nabla_x \cdot \mathbf{v}, \tag{1.18}$$

$$\nabla_x^2 \Phi = 4\pi G \rho. \tag{1.19}$$

The first of these is Euler's equation, which relates the velocity field $\mathbf{v}$ to the density field, pressure gradient and the gravitational potential Φ; the second is simply a statement that mass (or energy) is conserved and the third is Poisson's equation for the gravitational potential. In these equations $\mathrm{D}/\mathrm{D}t \equiv \partial/\partial t + \mathbf{v} \cdot \nabla$ is the convective derivative and I use ∇_x to signify taking a derivative with respect to the proper coordinates $\mathbf{x}$. The velocity field can be decomposed into a velocity due to the Hubble flow and a peculiar velocity: $\mathbf{v} = H\mathbf{x} + R\mathbf{u}$. It is more convenient at this point to work in comoving coordinates $\mathbf{r} = \mathbf{x}/R$. Defining $\delta \equiv \delta\rho/\rho_0$ to be the fractional overdensity, equations (1.17) and (1.18) can be linearised by keeping only first-order terms in the perturbed quantities (δ, $\mathbf{u}$, etc.), to give

$$\dot{\mathbf{u}} + 2\frac{\dot{R}}{R}\mathbf{u} = -\frac{1}{R^2}\nabla_r \delta\Phi - \frac{1}{R^2\rho_0}\nabla_r \delta p, \tag{1.20}$$

$$\dot{\delta} = -\nabla_r \cdot \mathbf{u}, \tag{1.21}$$

where ∇_r is the spatial derivative with respect to comoving coordinates. For a matter-dominated universe, the pressure term in equation (1.20) can be ignored, since non-relativistic matter is essentially pressureless. Equations (1.20) and (1.21) can be combined with (1.19) to give an expression for the growth of linear perturbations of comoving wavenumber k:

$$\ddot{\delta} + 2\frac{\dot{R}}{R}\dot{\delta} - 4\pi G\rho_0 \delta = 0. \tag{1.22}$$

In general, there are two solutions, $D_1(t)$ and $D_2(t)$, to this equation. Because the equation is linear, we can write a general solution of the form

$$\delta(t) = A_1 D_1(t) + A_2 D_2(t). \tag{1.23}$$

The behaviour of a matter-dominated, flat ($\Omega = 1$) universe is particularly simple: I have already shown (section 1.1) that, for this simple case, the scale factor scales as $R(t) \propto (t/t_0)^{2/3}$, where $t_0 = 2/(3H_0)$ and it follows (using equation 1.6) that

$$4\pi G\rho_0(t) = \frac{2}{3}\left(\frac{t}{t_0}\right)^{-2}. \tag{1.24}$$

It can easily be shown, by substitution, that the two solutions to equation (1.22) are

$$D_1(t) \propto t^{2/3}, \tag{1.25}$$

$$D_2(t) \propto t^{-1}. \tag{1.26}$$

The first of these is the growing mode, while the second decays rapidly, so that the growth of perturbations in the linear regime is described by

$$\delta(t) = \delta(t_0)\frac{D_1(t)}{D_1(t_0)}. \tag{1.27}$$

For this reason, $D_1(t)$ is often called the linear growth factor.

During the radiation-dominated era, or in universes with vacuum energy, the pressure term in equation (1.20) is important. For adiabatic perturbations, the pressure is related to the density by

$$\delta p = \frac{\partial p}{\partial \rho}\delta\rho = c_s^2 \rho_0 \delta, \tag{1.28}$$

where c_s is the sound speed. If we consider plane wave solutions of the form $\delta(\mathbf{r}) \propto \mathrm{e}^{-i\mathbf{k}\cdot\mathbf{r}}$ then equation (1.22) becomes

$$\ddot{\delta} + 2\frac{\dot{R}}{R}\dot{\delta} - \left(4\pi G\rho_0 - \frac{c_s^2 k^2}{R^2}\right)\delta = 0, \tag{1.29}$$

for perturbations of comoving wavenumber k. For a radiation-dominated universe this yields $D_1(t) \propto t, D_2(t) \propto t^{-1}$. For universes with vacuum energy, the form of $R(t)$ changes and there is no general analytic solution. Fig. 1.3 illustrates the linear growth rate of perturbations for different cosmological models.

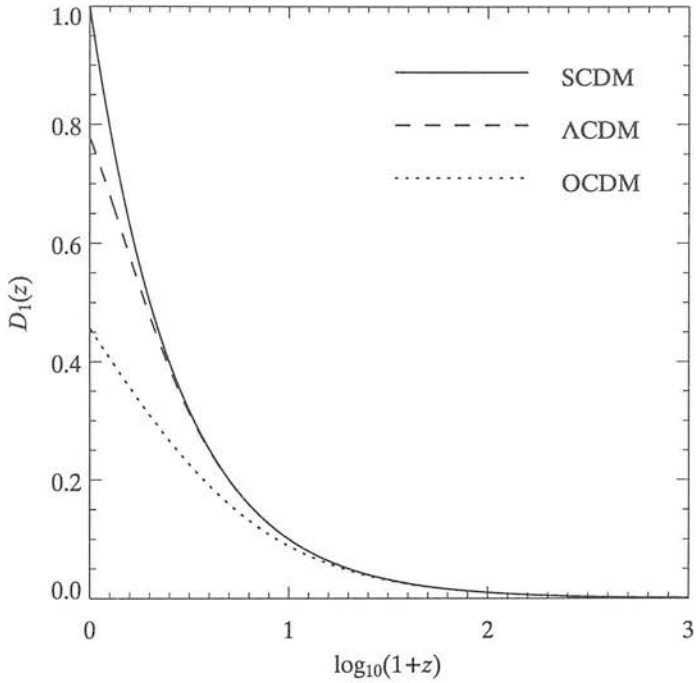

Figure 1.3: Evolution of the linear growth factor with cosmic epoch for the same three models as Fig. 1.1. The curve for each model is normalised to unity for the flat, matter-dominated model (solid line) at $z = 0$. Growth is suppressed in the vacuum-dominated model (dashed line), but not by as much as in the open model (dotted line).

1.2.3 Non-linear spherical collapse

The linear approximation used in section 1.2.2 is valid for small overdensities ($\delta \ll 1$); when $\delta \sim 1$, the non-linear terms in equations (1.17) – (1.19) become dominant and the simple treatment of the previous section is no longer adequate. The formation of galaxy haloes — a highly non-linear process — can be understood by considering the simple model of a uniform, spherical overdensity that will eventually form a virialised object of mass M. An overdense sphere behaves exactly like a closed universe, that is, its motion is governed by analogues of the Friedmann equations (equations 1.1 and 1.2):

$$\dot{r}^2 - \frac{8\pi G\rho}{3} r^2 = \text{constant} \tag{1.30}$$

$$\ddot{r} = -\frac{4\pi G\rho}{3} r. \tag{1.31}$$

It can be shown that the general solution (in parametric form) to equation (1.30) is

$$r(\theta) = A(1 - \cos\theta) \tag{1.32}$$

$$t(\theta) = B(\theta - \sin\theta), \tag{1.33}$$

where $A^3 = GMB^2$. It can be seen immediately that $r(\theta)$ has a maximum at $\theta = \pi$; at this point the protogalaxy breaks away from the general expansion of the universe and begins to collapse under its own self-gravity. A completely uniform, spherically-symmetric perturbation will eventually collapse to a single point at $\theta = 2\pi$. In practice, however, the existence of substructure leads to chaotic, time-varying gravitational potentials which have the effect of allowing particles to exchange energy, thereby converting the ordered motion of the collapse into random motions in a process known as violent relaxation (Lynden-Bell 1967). At $\theta = 3\pi/2$, the kinetic energy of the collapse is equal to exactly one half of the gravitational potential energy, the condition for virialisation. Using equation (1.32) we can calculate the mean density of the perturbation at virialisation:

$$\rho_{\rm vir} = \frac{3M}{4\pi r_{\rm vir}^3}, \tag{1.34}$$

where $r_{\rm vir} = r(\theta = 3\pi/2)$ is the virial radius. A more useful quantity is the mean overdensity, $1 + \delta_{\rm vir} \equiv \rho_{\rm vir}/\rho$, at virialisation. To calculate this, we need to know the mean density of the universe at the time when the perturbation achieved virialisation.

Considering once more the simple model of a matter-dominated, flat universe, the mean density at time t is given by equation (1.24). If virialisation indeed occurs at the expected time, given by equation (1.33) with $\theta = 3\pi/2$, then combining equations (1.34) and (1.24) gives the overdensity as $1 + \delta_{\rm vir} \simeq 147$. A common assumption is that virialisation is not completed until the time at which the perturbation would have collapsed to a single point; in this case the overdensity is slightly higher: $1 + \delta_{\rm vir} \simeq 178$. Another quantity often quoted is the extrapolated linear overdensity, $\delta_{\rm lin} \propto t^{2/3}$, at virialisation: $\delta_{\rm lin} = 1.69$ (or 1.06 if virialisation occurs at $\theta = 3\pi/2$).

The above derivation relies on the assumption that the universe is matter-dominated with a flat geometry; for a flat universe with a cosmological constant, the appropriate value can be shown (White, Efstathiou & Frenk 1993) to be

$$1 + \delta_{\rm vir} \simeq 178\Omega_{\rm M}^{-0.6}. \tag{1.35}$$

The importance of this result is that it allows us to define exactly what we mean by a galaxy halo: a region of space within which the average overdensity is given by equation (1.35). Numerical simulations (e.g. Cole & Lacey 1996) show that the virial radius, $r_{\rm vir}$, which defines the extent of the halo, delineates well the transition from chaotic motions to ordered inflow of material from the immediate surroundings of the halo.

In practice, since the exact time of virialisation is uncertain, a value of $1+\delta_{\rm vir} \simeq 200$ is often assumed. Given this uncertainty, the dependence on $\Omega_{\rm M}$ is only a small effect (a factor of ~ 2 for $\Omega_{\rm M} = 0.3$). In what follows, I use the common notation r_{200}, V_{200}, etc. to denote virial quantities (those measured at, or within, the virial radius) predicted by the spherical collapse model, but I use the exact expression (equation 1.35) in the calculations.

The mean density of a halo with mass M and virial radius r_{200} is

$$\bar{\rho}(r < r_{200}) = \frac{3M}{4\pi r_{200}^3}. \tag{1.36}$$

It is often more convenient to specify a halo by its virial velocity,

$$V_{200} = \left(\frac{GM}{r_{200}}\right)^{1/2}, \tag{1.37}$$

rather than its virial radius. The average density is then given by

$$\bar{\rho}(r < r_{200}) = \frac{3V_{200}^2}{4\pi G r_{200}^2}. \tag{1.38}$$

Setting this equal to $178\Omega_{\rm M}^{-0.6}\rho_{\rm c}$ and using equation (1.37) gives useful relationships between the mass of a halo, its circular velocity and its virial radius:

$$V_{200} = [9.4\Omega_{\rm M}^{-0.3} GMH(z)]^{1/3} \tag{1.39}$$

$$r_{200} = \frac{GM}{V_{200}^2} \tag{1.40}$$

$$= \frac{V_{200}}{9.4\Omega_{\rm M}^{-0.3} H(z)}. \tag{1.41}$$

Thus, the mean density, and hence the circular velocity, of a halo of a given mass depends strongly on the cosmological model adopted, mostly as a result of the differing behaviour of $H(z)$ in different models of the universe, as illustrated in Fig. 1.1.

1.2.4 The formation of disc galaxies

White & Rees (1978) set out the now standard picture for the formation of galaxies. During the collapse and virialisation of a dark matter halo, the baryonic component is shock-heated to the virial temperature of the halo. At the temperatures typical of galaxy haloes, the gas is fully ionised and can dissipate energy efficiently by means of bremsstrahlung radiation, so it will continue to collapse until it reaches centrifugal equilibrium. The first detailed description of this process is usually credited to Fall & Efstathiou (1980). In their model, a galaxy begins life as a uniformly rotating, homogeneous cloud of gas. The initial angular momentum of the protogalaxy is built up through tidal torques from the surrounding structures before the galaxy condenses out of the Hubble flow (Hoyle 1953; Peebles 1969). At this point, the baryons closely follow the dark matter and hence experience the same tidal field and are expected to have the same specific angular momentum; the two components do not separate out until after virialisation.

If the collapse of the gas is smooth then angular momentum is conserved and the gas forms a thin, rapidly rotating disc, whose size is determined by the balance between gravity and the centrifugal force. Peebles (1971) introduced a dimensionless spin parameter,

$$\lambda \equiv \frac{J|E|^{1/2}}{GM^{5/2}}, \tag{1.42}$$

to describe the angular momentum of a protogalaxy of mass M and angular momentum J, with binding energy E. Dark matter haloes are expected to have $\lambda_{\rm med} \simeq 0.04$ (e.g. Cole & Lacey 1996). For a flat rotation curve the specific angular momentum scales linearly with radius, so to reach centrifugal equilibrium ($\lambda = 1$) the gas cloud has to collapse by a factor of order $1/\lambda \sim 25$. If the origin of the angular momentum is in tidal torques then it is expected that the material with the lowest angular momentum will be closest to the centre of the halo and so will be the first to collapse (because the higher density makes cooling more efficient), leading to 'inside-out' disc formation.

The idea of a homogeneous, uniformly rotating protogalaxy is clearly at odds with the hierarchical model of structure formation, which predicts that galaxies formed by merging of smaller sub-units. Shocks caused by substructure in the dark matter

distribution are likely to compress the gas, allowing it to dissipate energy and condense into clumps. The process of building up a galaxy by mergers between smaller gaseous subclumps is efficient at transporting angular momentum outwards into the halo; this behaviour is seen in gas-dynamical simulations (Navarro, Frenk & White 1995) and is often called the *angular momentum catastrophe*, since it results in discs that are much smaller than those of real galaxies. This a serious problem for hierarchical models. One way to prevent the catastrophic cooling of gas into substructure is to postulate strong feedback from star formation (White & Rees 1978; White & Frenk 1991), which has the effect of reheating the gas and maintaining a smooth distribution until after the majority of the substructure has been erased. Any stars formed in subclumps will have their orbits mixed collisionlessly to form an old spheroidal component to the galaxy. An argument for such early star formation was put forward by Eggen, Lynden-Bell & Sandage (1962) on the basis that low metallicity stars in the solar neighbourhood are observed to have highly eccentric orbits, compared to those of solar metallicity stars, implying that stars in the Galactic spheroid formed on a shorter timescale.

Light profiles

The surface brightness profiles of spiral galaxies are well-fitted by an exponential law:

$$I(R) = I_0 \mathrm{e}^{-R/R_\mathrm{d}} \tag{1.43}$$

in the outer parts, where the thin, disc component dominates over the spheroid. This is often taken to imply that the underlying mass distribution is also exponential, although this is not necessarily the case. If angular momentum transport during collapse is not important, then the arrangement of material in discs reflects the initial angular momentum distribution of gas in haloes. Another possibility, put forward by Lin & Pringle (1987), is that the exponential profile does not reflect the distribution of gas falling onto the disc but is the result of evolution. They showed that if viscosity is effective at transporting angular momentum outwards in discs then exponential *stellar* profiles occur naturally, provided that angular momentum transport acts on a timescale of the same order as the star formation timescale. This is true even when the initial gas profile is much flatter.

ROTATION CURVES

The rotation curves of disc galaxies are remarkably flat at large radii, a property that extends well beyond the optical galaxy, in the larger neutral gas disc. This, of course, is part of the evidence for the existence of dark matter in galaxies. The outer regions of discs contain little mass; the dynamics should be dominated by the halo, so whatever profile the dark matter has must be responsible for the flat rotation curves.

1.2.5 The formation of ellipticals and bulges

The spheroidal components of galaxies, both ellipticals and the bulges of spiral galaxies, are not naturally included in the Fall & Efstathiou model of galaxy formation. They have very little angular momentum and are almost entirely supported by the 'pressure' of radial stellar orbits. One possible explanation for their kinematical properties is that the spheroid stars formed early in the collapse of the protogalaxy and that their subsequent collapse was dissipationless. This is the Eggen, Lynden-Bell and Sandage picture and, while it may explain the old stellar halo of our own Galaxy, the high densities of bulges requires significant dissipation. Alternatively, one can appeal to some mechanism by which angular momentum is removed from the central regions of galaxies.

In the hierarchical scenario, spheroids naturally form in mergers between disc galaxies, as postulated by Fall (1979). This currently seems the most promising scenario: N-body studies (e.g. Barnes 1988; Hernquist 1992, 1993; Heyl, Hernquist & Spergel 1994) have shown that mergers between equal-mass disc galaxies produce remnants with elliptical isophotes and $R^{1/4}$–law surface brightness profiles. The high phase-space density of ellipticals requires significant dissipation (Hernquist, Spergel & Heyl 1993), i.e. a significant gaseous component. Gas-rich mergers result in strong flows of gas into the nuclei of the merging galaxies, fuelling a starburst (Mihos & Hernquist 1994a, b, 1996) which adds to the light of the spheroidal component.

Small bulges have properties more closely related to those of the disc in which they are embedded (Courteau, de Jong & Broeils 1996; Peletier & Balcells 1996), suggesting that secular evolution of the disc may also play a role in the formation of some bulges.

Zhang & Wyse (2000) posit that angular momentum transport through viscosity in the early disc may be responsible for the formation of small bulges. Bars are also efficient at transporting angular momentum away from the central regions of discs; they cause vertical heating of the disc and eventually dissolve to form peanut-shaped bulges (Combes et al. 1990).

Light profiles

The surface-brightness profiles of ellipticals and large spiral bulges are generally well-fit by de Vaucouleurs' (1948) $R^{1/4}$–law:

$$I(R) = I_{\rm e} {\rm e}^{-7.67[(R/R_{\rm e})^{1/4}-1]}, \tag{1.44}$$

while small bulges and dwarf ellipticals are better fit by an exponential profile (Graham 2001). The isophotes of elliptical galaxies are close to being ellipses; deviations from perfect ellipticity lead to elliptical galaxies being described as either 'boxy' (meaning that their isophotes have the appearance of having been compressed along the major and minor axes) or 'discy' (when their isophotes have been stretched out along these same axes). The shapes of the isophotes of elliptical galaxies are correlated with their luminosities — in the sense that bright ellipticals tend to have boxy isophotes while low-luminosity ellipticals and bulges tend to be discy — and with their kinematical properties (see below).

Kinematical properties

The kinematics of the bulges of early-type spirals are consistent with their being oblate, isotropic rotators — that is, their velocity dispersions are isotropic and they are flattened by rotation about their minor axes. The degree of rotational support in elliptical galaxies decreases with increasing luminosity, so that giant ellipticals are entirely pressure supported while low-luminosity ellipticals have the kinematical properties comparable to bulges of the same luminosity. The degree of rotational support is also correlated with isophote shape: ellipticals and bulges with a high degree of rotational support tend to have discy isophotes, while those of giant, slowly-rotating ellipticals are boxy.

Table 1.1: Recent measurements of the field galaxy luminosity function.

Survey	$\Phi^\star/10^{-2}h_0^3$ Mpc^{-3}	$M^\star - 5\log_{10} h_0$	α	Reference
B-band:				
2dFGRS[1]	1.68 ± 0.08	-19.40 ± 0.07	-1.21 ± 0.02	Norberg et al. (2002)
SDSS-comm[2,3]	2.06 ± 0.23	-19.51 ± 0.04	-1.26 ± 0.05	Blanton et al. (2001)
ESO Slice Project[1]	2.0 ± 0.4	$-19.35^{+0.06}_{-0.08}$	$-1.22^{+0.06}_{-0.07}$	Zucca et al. (1997)
Stromlo-APM[1]	1.40 ± 0.17	-19.24 ± 0.13	-0.97 ± 0.15	Loveday et al. (1992)
K-band:				
Hawaii-AAO	1.3 ± 0.3	-23.70 ± 0.08	-1.37 ± 0.10	Huang et al. (2003)
2dFGRS/2MASS[4]	1.08 ± 0.16	-23.44 ± 0.03	-0.96 ± 0.05	Cole et al. (2001)
2MASS/CfA2[4]	1.16 ± 0.1	-23.39 ± 0.05	-1.09 ± 0.06	Kochanek et al. (2001)
Stromlo-APM	1.2 ± 0.8	-23.58 ± 0.42	-1.16 ± 0.19	Loveday (2000)
Autofib-2	1.66	-23.12 ± 0.04	-0.91 ± 0.04	Gardner et al. (1997)

[1] Converted from b_J as described in the text

[2] Converted from SDSS $g^\star$ magnitudes as described in the text

[3] A re-analysis by Norberg et al. yields $\Phi^\star = 1.56 \times 10^{-2}h_0^3$ Mpc^{-3}, $M^\star = -19.42 + 5\log_{10} h_0$ and $\alpha = -1.26$ (see text for details)

[4] Uses 2MASS Kron magnitudes (K_S)

1.3 The galaxy population at $z = 0$

1.3.1 Luminosity function of galaxies

The most fundamental observable property of the galaxy population is the luminosity function $\Phi(L)$ — the comoving space density of galaxies of a given luminosity over some range of wavelengths. The shape of the luminosity function can be conveniently parameterised by a Schechter function (Schechter 1976):

$$\Phi(L)\,\mathrm{d}L = \Phi^\star \left(\frac{L}{L^\star}\right)^\alpha \exp\left(-\frac{L}{L^\star}\right) \frac{\mathrm{d}L}{L^\star}. \tag{1.45}$$

This has the properties that at low luminosities it rises (or falls) monotonically with decreasing luminosity, with a slope α, while above the characteristic luminosity, $L^\star$ (often quoted as the equivalent magnitude, $M^\star$), it turns over and falls away abruptly. Although there are many more faint galaxies than bright ones, the luminosity density of the universe, $L\phi(L)$ is dominated (at least at optical wavelengths) by galaxies of around the characteristic luminosity $L^\star$. In Fig. 1.4 I plot recent measurements of the luminosity function in two optical bands; the best-fitting Schechter function parameters

are listed in table 1.1. Data points are plotted for the two most recent surveys in each band: in the B band from the *2dF Galaxy Redshift Survey* (2dFGRS; Norberg et al. 2002) and the *Sloan Digital Sky Survey* (SDSS; Blanton et al. 2001); and in the K band the combined 2dFGRS/2MASS data (Cole et al. 2001) and the Hawaii-AAO survey (Huang et al. 2003). In each case, the Schechter function representation fits the data points extremely well over more than three orders of magnitude in luminosity. Magnitudes measured in the $b_{\rm J}$ band have been converted to the standard Johnson B band using the colour equations of Blair & Gilmore (1982):

$$B = b_{\rm J} + 0.28(B - V). \tag{1.46}$$

The B-band luminosity function of the SDSS commissioning data has been estimated from that measured in the SDSS $g^\star$ band using the colour equation of Fukugita et al. (1996):

$$B = g^\star + 0.44(B - V) + 0.12. \tag{1.47}$$

In both cases I have assumed a mean galaxy colour of $\langle B - V \rangle = 0.94$ (Norberg et al. 2002). In spite of the good fit to the data in each independent measurement of the luminosity function, the parameters of the Schechter function required to fit the data vary widely from survey to survey and the shape of the luminosity function has been shown to vary with galaxy colour, environment, spectral type (Madgwick et al. 2002; Folkes et al. 1999) and morphology (Marzke et al. 1998). Values of $M^\star$ derived from surveys in a given band are consistent, but the overall normalisation varies by almost a factor of 2 between the surveys listed in table 1.1. However, Norberg et al. have re-analysed the Blanton et al. data and find that if the same method is used to normalise the Schechter functions and the correct colour equation (1.46) is used then the parameters measured from the SDSS and 2dFGRS surveys agree to within the quoted errors. The least well constrained parameter is the faint-end slope, α, with different surveys failing even to agree on whether the number of galaxies increases or decreases towards fainter magnitudes. The most recent surveys in the $b_{\rm J}$-band (Zucca et al. 1997; Norberg et al. 2002) appear to be converging on a steep faint-end slope. The faint end of the K-band luminosity function is less accurately measured but, while there is still considerable disagreement over the slope, most of the measurements listed

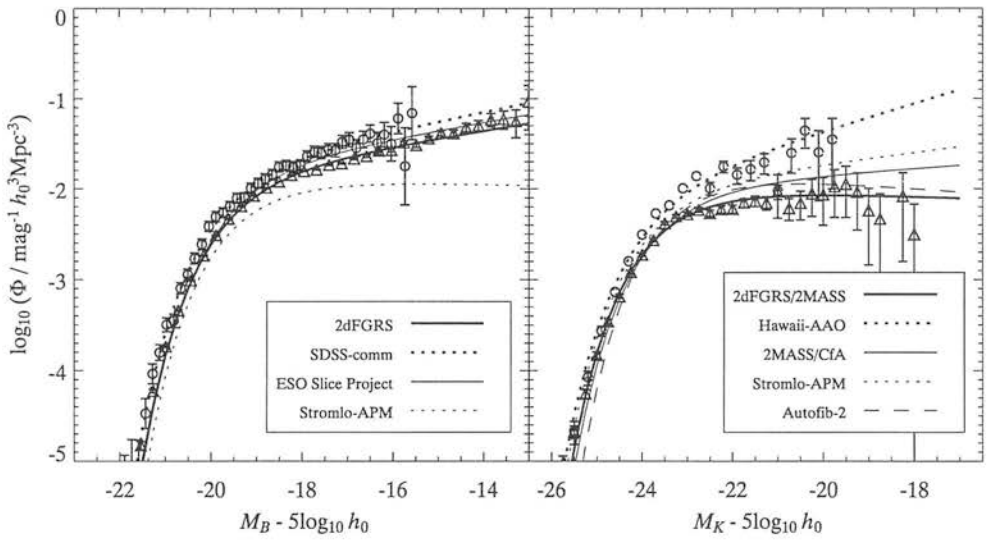

Figure 1.4: A compilation of recent measurements of the field galaxy luminosity functions in the B and K bands. The various curves are Schechter function parameterisations of the results of different surveys, as described in table 1.1. The most recent results are plotted as data points: in the B band from Norberg et al. (2002, triangles) and Blanton et al. (2001, circles); and in the K-band from Cole et al. (2001, triangles) and Huang et al. (2003, circles).

are consistent with $\alpha = -1$.

A large part of the discrepancy between measurements of the luminosity function is likely to be the result of the differing selection criteria of the galaxy surveys used. The discovery of significant numbers of low surface-brightness (LSB) galaxies in the local universe (Schombert et al. 1992; O'Neil, Bothun & Cornell 1997) has led to the suggestion that existing galaxy surveys are missing large numbers of galaxies due to their surface brightness limits. Including these missing galaxies may lead to a considerable change in the derived luminosity function and may resolve the discrepancies between the different measurements of the faint-end slope given in table 1.1. Bothun, Impey & Malin (1991) have investigated the effects of surface brightness limits on the luminosity functions of clusters, where LSB galaxies are most easily identified in large numbers. They measure faint-end slopes of $\alpha = -1.5$ for the Fornax cluster and $\alpha = -1.6$ for the Virgo cluster (using the sample of Impey, Bothun & Malin 1988), compared to $\alpha = -1.32$ derived by Ferguson & Sandage (1988) for the Fornax cluster,

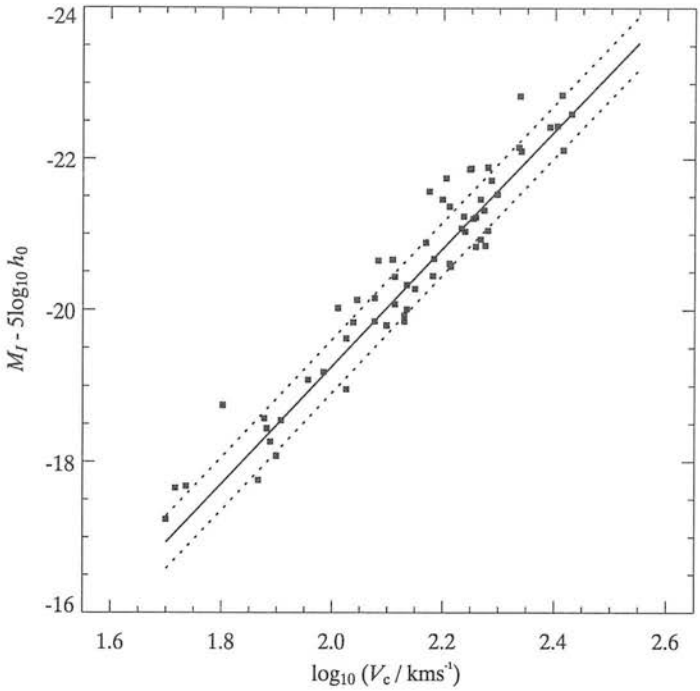

Figure 1.5: The infrared Tully–Fisher relation for spiral galaxies. The solid line is the mean of four measurements of the I-band Tully–Fisher relation (Pierce & Tully 1992, Mould et al. 1993, Giovanelli et al. 1997, Mathewson, Ford & Buchhorn 1992) and the dotted lines show the typical scatter in the observed relation (Giovanelli et al. 1997). The data points are for the sample of Tully et al. (1998) and are plotted to illustrate the scatter around the mean relation.

from a shallower survey. No similar study exists for field galaxies, but Sprayberry et al. (1997) have measured the luminosity function of field LSB galaxies and find their space density is comparable to, or greater than, that of high surface brightness galaxies at magnitudes $M_B \gtrsim -15$. They also find evidence for a steepening of the luminosity function of LSBGs at $M_B \gtrsim -16$. This suggests that existing galaxy surveys *are* missing a significant population of LSB galaxies and that these contribute strongly to the luminosity function at faint magnitudes. Luminous LSB galaxies, however, are rare so that these galaxies cannot significantly alter measurements of the total luminosity density of the universe, which is dominated by $L^\star$ galaxies.

1.3.2 The infrared Tully–Fisher relation

The luminosities of spiral galaxies in many wavelength ranges are strongly correlated with their rotational velocities (Tully & Fisher 1977). The rotational velocity is usually related to the velocity-width of HI atomic emission lines, corrected for the inclination of the disc. In Fig. 1.5 I plot the mean of four measurements of the Tully-Fisher relation in the I band. Working at infrared wavelengths has several advantages. Firstly, corrections have to be applied to the measured luminosities to account for dust extinction, both internal and within our own galaxy; these are much smaller at infrared wavelengths than in the optical, so that the results have smaller uncertainties. Infrared light is also a better tracer of the total stellar mass than bluer wavelengths, which are strongly affected by small amounts of recent star formation. The amount of internal dust extinction also depends on inclination; commonly, luminosities are corrected to face-on, dust-extincted values. The scatter about the mean relation in the I band for a given sample of galaxies is typically 0.35 magnitudes, although the scatter between different samples is often much larger. The latter is probably caused by differences in the methods used to measure and correct the luminosities and circular velocities of galaxies.

The Tully–Fisher relation is principally a relationship between dynamical mass of a galaxy (as probed by the rotational velocity of the disc, and including dark matter) and the luminous mass (traced by the luminosity at some wavelength). The Tully–Fisher relation therefore constrains the mass-to-light ratios of spiral galaxies. Low surface-brightness galaxies are generally found to lie on the same Tully–Fisher relation as their high surface-brightness counterparts, implying total mass-to-light ratios that are ~ 2 times higher (Zwaan et al. 1995). However, O'Neil, Bothun & Schombert (2000) give several examples of LSB galaxies that appear to be underluminous for their HI line-widths, so the situation is at least a little unclear.

1.3.3 The surface-brightness function

Freeman (1970) found that the central surface brightnesses of spiral galaxy disks appeared to have a Gaussian distribution centred on $\mu_0 = 21.65$ $B-$mag arcsec^{-2} and with a very narrow width (0.3 $B-$mag arcsec^{-2}). Had this turned out to be an in-

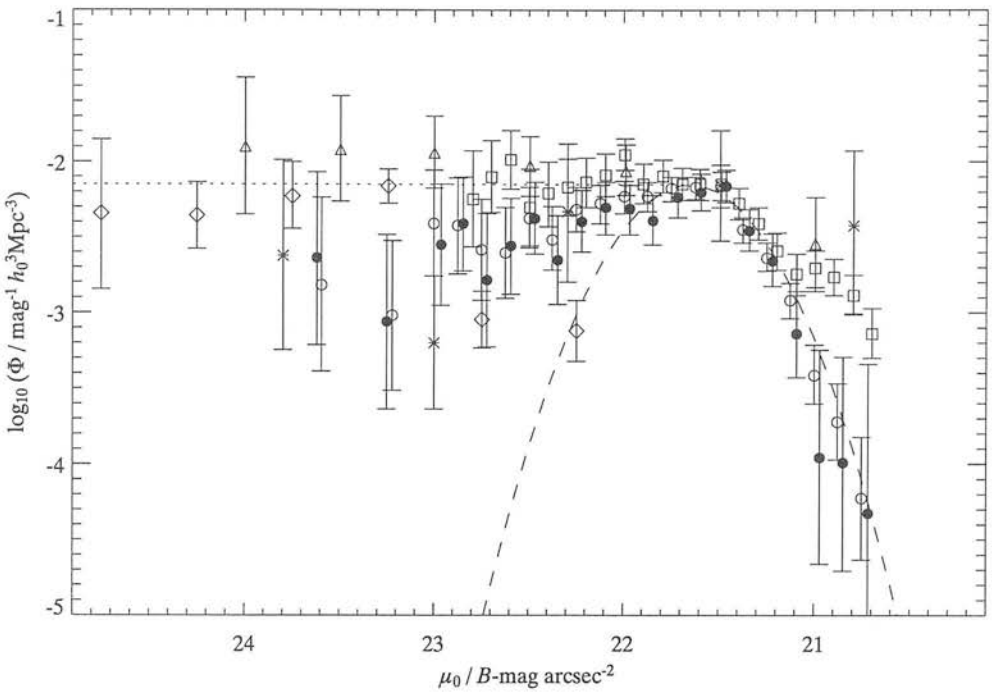

Figure 1.6: Number density of galaxies as a function of disk central surface brightness. The data shown was compiled by O'Neil & Bothun (2000) from the following sources: O'Neil, Bothun & Schombert (2000, open diamonds), Phillipps et al. (1987, open squares), Davies (1990, circles: filled for a flux-limited sample and open for diameter-limited), Sprayberry (1994, open triangles) and de Jong (1996, stars). The dashed curve is the Gaussian distribution found by Freeman (1970), while the data are consistent with a flat distribution (dotted line) down to the limits of current surveys.

trinsic property it would have placed severe constraints on models of galaxy formation: properties such as stellar mass, mass-to-light ratio (which is determined by the star formation history) and angular momentum must all conspire to give a constant surface brightness. It was first pointed out by Disney (1976) that the rapid fall-off in the numbers of galaxies observed at surface brightnesses approaching that of the night sky could result from selection effects. Since then, the existence of a large population of LSB galaxies has been confirmed. Taking surface brightness selection effects properly into account, O'Neil & Bothun (2000) have shown that current observations are consistent with a constant number density of galaxies per magnitude interval down to 25 $B-$mag arcsec^{-2} and beyond, as shown if Fig. 1.6. These galaxies are below the

detection thresholds of most large galaxy surveys, even though their integrated fluxes often place them well inside the formal magnitude limits.

1.4 Summary and important questions

I have now set out the basic picture of galaxy formation on which the remainder of this thesis is based. In this picture, the universe is dominated by some dark form of matter. Small fluctuations in the matter density field collapse under gravity to form the potential wells into which the baryons fall after recombination. The smallest of these structures reach virialisation first and are susequently incorporated into larger structures in a process of hierarchical structure formation. After virialisation, the baryons dissipate energy through radiation and collapse to form thin discs in centrifugal equilibrium. Such gas discs are the main sites of star formation and feedback from this process may reheat some of the gas to the hot phase. The most likely mechanism for the formation of ellipticals and large bulges is in mergers between disc galaxies, which are a natural consequence of the hierarchical scenario.

This picture has been successful at explaining many of the key properties of the galaxy population, but there are a number of important questions that remain unanswered. Many of these are related to the overabundance of low-mass haloes and subhaloes predicted by the CDM model:

1. *What determines the faint-end slope of the luminosity function?* Naive models of galaxy formation, combined with the CDM model of structure formation, predict a steep faint-end slope, which conflicts with observations. Feedback from star formation may provide the answer, but a high efficiency is required and there is little observational evidence of its importance in normal galaxies.

2. *How much of the initial angular momentum is lost during the galaxy formation process?* Simulations suggest that a significant fraction is lost, but this is inconsistent with the observed sizes of disc galaxies. The 'angular momentum catastrophe' is caused by gas cooling into clumps in subhaloes, which then lose angular momentum through dynamical friction. Again, stellar feedback may be a

solution, preventing gas from cooling into small units and maintaining a smooth distribution until galaxy-scale haloes form.

3. *How fragile are galaxy discs?* Discs are generally thought of as fragile structures which should, at the very least, be heated and thickened by the substructure passing through them. This is the case even if such substructure is unable to host visible galaxies because of strong feedback, so feedback is not a solution in this case.

More general questions include:

1. *What determines the luminosities of the brightest galaxies?* Explaining the exponential cut-off in the luminosity function above $L^\star$ requires a mechanism capable of shutting off cooling in massive galaxy clusters or, at least, of preventing any gas that does cool from forming luminous stars. Several possiblities are commonly considered; these include heat conduction in the gas and reheating, possibly by active galactic nuclei (AGN).

2. *What is the significance of the low-surface brightness (LSB) galaxy population?* LSB galaxies are difficult to observe and so their numbers are uncertain; although they are believed to make only a small contribution to the local luminosity density of the universe, they could, in principle, contain a significant fraction of the baryons. Any viable model for galaxy formation must, in any case, be able to predict their numbers and properties.

3. *What role do AGN play in galaxy formation?* So far I have not mentioned AGN (with the exception of the possibility of their suppressing cooling in clusters). They are generally ignored in galaxy formation models, but it has been estimated that their combined energy output, at least at optical wavelengths, is comparable to that from stars, so they might be expected to have played some role in galaxy formation.

I will not attempt to answer all of these questions, but I will put forward potential solutions to some of them, including alternatives to the solutions already discussed.

The remainder of this thesis is laid out as follows. In chapter 2 I describe a phenomenological approach to modelling galaxy formation, beginning with a method for following the formation and merging of galaxy haloes in an N-body simulation. I then outline my prescriptions for modelling radiative cooling, star formation and the associated feedback processes and the merging of satellite galaxies within haloes and for the evolution of stellar populations. The free parameters of the model are chosen to give a good fit to the luminosity functions at two optical wavelengths and to the infrared Tully–Fisher relation. In chapter 3 I build on the basic model by adding a more detailed description of galaxy discs, including an observationally-motivated star formation law with a surface density threshold. I investigate the effect of such a threshold not only on the properties of galaxy discs but also on the global properties of galaxies. In chapter 4 I apply the model developed in the preceding two chapters to the question of the role played by selection effects in shaping the luminosity function. I construct magnitude-limited mock surveys with a range of isophotal limits and then attempt to reconstruct the luminosity function using a simple version of one of the techniques used by observers. Finally, in chapter 5, I discuss the limitations of my current model and present some preliminary findings from a more advanced disc model. I suggest a number of possible further improvements and end by discussing the potential for future work in this area.

Chapter 2

Overmerging in galaxy formation models

In this chapter I describe a phenomenological approach to modelling galaxy formation in N-body simulations. I begin by discussing the problem of overmerging in galaxy formation models and I outline a method for following the formation and merging of galaxy haloes in N-body simulations, which is optimised to find and preserve substructure. Groups of particles representing collapsed, virialised objects (haloes) are identified using local density percolation and replaced by single softened particles (van Kampen 1995) to prevent disruption of such groups in high density regions by 2-body heating and artificial tidal stripping. This galaxy formation algorithm is applied many times during the evolution of the simulation, to build up a merger tree for each halo. I then set out the basic ingredients of a more detailed model for the formation of galaxies within dark matter haloes: radiative cooling of gas in haloes, star formation and the associated feedback processes, the merging of galaxies within merged haloes and, finally, stellar spectral synthesis, needed to predict the visible properties of galaxies. The model presented here is similar in many ways to the popular technique of semi-analytic modelling, but the coupling of a detailed model for galaxy formation with an N-body simulation incorporating a built-in recipe for the formation of galaxy haloes makes this work unique and represents an improvement over earlier models. In particular, it provides a natural explanation for the exponential cut-off at the bright end

of the galaxy luminosity function — a problem for many galaxy formation models — without the addition of new baryonic physics.

2.1 Introduction

Cole et al. (1994) identify four key ingredients necessary to any complete model of galaxy formation:

1. The evolution of the dark matter distribution in its proper cosmological setting, including the non-linear processes associated with the collapse and merging of dark matter haloes.

2. The dynamical behaviour of gas coupled gravitationally to the dark matter and subject to shocks, cooling and heating and processes.

3. The evolution of stellar populations, including their spectrophotometric properties as a function of time and the injection of energy, mass and metals into the surrounding interstellar and intergalactic gas.

4. Tidal interactions and mergers of individual galaxies in a dynamically active environment.

The first of these is the most studied and by far the best understood; the behaviour of dark matter on large scales can be modelled relatively easily since the only important force is gravity. Modelling the collapse and virialisation of dark matter haloes exactly requires the use of numerical techniques, such as N-body simulations, because it involves highly non-linear processes. In spite of this, analytic approaches based on an extension of the Press-Schechter formalism (Bond et al. 1991; Bower 1991) have been remarkably successful. Such methods have the advantages of high mass and time resolution, without the need for powerful computing resources, as demanded by the more direct, numerical approaches.

Gas dynamics can also be treated numerically, the most common technique used for such purposes being smoothed-particle hydrodynamics (SPH; Gingold & Monaghan 1977), in which smoothed particles are used to represent fluid elements. Radiative

cooling can be taken into account using a cooling function (e.g. Katz & Gunn 1991). Because of the similarities between SPH and N-body simulations — both use discrete particles to sample a continuous density field — the two are easily combined to create hybrid codes capable of following the evolution of both dark matter and baryons, and SPH has found widespread use in cosmology for this reason. Both of the above techniques have been employed extensively to study galaxy–galaxy mergers and interactions (collisionless N-body by Barnes 1992; Walker, Mihos & Hernquist 1996; Naab, Burkert & Hernquist 1999; Naab & Burkert 2001a and gas-dynamical by Mihos & Hernquist 1994a, b, 1996; Barnes & Hernquist 1996; Naab & Burkert 2001b).

Compared to the other ingredients, the physics of star formation and, in particular, its relationship to galaxy formation is not well understood. No firm theoretical predictions exist for the rate of star formation in galaxies, so galaxy formation models are forced to rely on either empirical rules (such as the observed law of Schmidt 1959) or simple scaling relations. Furthermore, it is not known whether the initial mass function (IMF) has a universal form or changes with environment or cosmic epoch. The evolution of individual stars is better understood (although some controversy still exists over the later stages of stellar evolution, as discussed by Charlot, Worthey & Bressan 1996). If one assumes a form for the IMF then it is possible to predict the properties of entire coeval populations. In particular, there exist libraries of synthetic stellar spectra (e.g. Bruzual & Charlot 1993; Worthey 1994; Jimenez et al. 1998) for stellar populations of different ages and metallicities, which can be used to relate the photometric properties of galaxies to their star formation history.

The technique of semi-analytical modelling, pioneered by White & Frenk (1991), was the first concerted attempt to bring together all of the ingredients for galaxy formation in a single model. In their most common form, semi-analytical models (SAMs) use Monte–Carlo realisations of the extended Press–Schechter formalism to obtain statistical distributions of halo properties and merger histories. The remaining ingredients are modelled using, where possible, simple physical laws or, where the physics is poorly understood, scaling relations. Using a SAM, White & Frenk were able to predict many of the basic properties of galaxies, including luminosities, circular velocities, metallicities and star formation rates. The semi-analytical approach was rapidly taken up,

Table 2.1: The cosmological model.

Ω_Λ	Ω_M	h_0	σ_8	Ω_b
0.7	0.3	0.7	0.93	0.04

extended and improved by many other groups, most notably Cole et al. (1994), Cole et al. (2000, CLBF), Kauffmann & White (1993), Kauffmann, White & Guiderdoni (1993), Kauffmann (1996) and, more recently, Somerville & Primack (1999). These models have successfully reproduced many of the global properties of galaxies in a hierarchical universe and have offered explanations of, among other things, the shape of the field galaxy luminosity function and the slope (although not the zero–point) of the Tully–Fisher relation.

An alternative to semi-analytical modelling is to identify galaxy haloes at a series of times in an N-body simulation and in this way to follow the merger histories of present-day galaxy haloes. This was the approach adopted by van Kampen, Jimenez & Peacock (1999, vKJP). Using an N-body simulation with a built in recipe for galaxy formation (van Kampen 1995), they found that it was possible to avoid the over-merging of haloes which is a limitation of the Press-Schechter formalism. Over-merging is the failure to identify or account for substructure within overdense regions. Preserving this substructure allowed vKJP to simultaneously fit the K-band luminosity function and the zero–point of the I-band Tully–Fisher relation. The idea of extracting merger trees from N-body simulations has recently been applied by several other groups (Helly et al. 2003; Kauffmann et al. 1999).

Throughout this thesis, I adopt the now standard ΛCDM cosmology. The parameters of the cosmological model are summarised in table 2.1. In addition to assuming a flat universe with $\Omega_M = 0.3$, I take a Hubble constant of $h_0 = 0.7$ (the Hubble Space Telescope Key Project to measure the Hubble constant yielded a value of 0.72 ± 0.08; Freedman et al. 2001) and a baryon density of $\Omega_b = 0.04$, consistent with the constraints from nucleosynthesis ($\Omega_b h_0^2 = 0.0190 \pm 0.0018$; Burles & Tytler 1998) and from the microwave background ($\Omega_b h_0^2 = 0.021$; Krauss 2003). I adopt a power spectrum

normalisation of $\sigma_8 = 0.93$ (Eke, Cole & Frenk 1996); this was originally chosen to be consistent with the constraints from the amplitude of CMB fluctuations as measured by COBE (Tegmark 1996; see Cole et al. 1997) but is now generally considered to be an overestimate. More recent estimates (e.g. Lahav et al. 2002) favour values of $\sigma_8 = 0.7 - 0.8$.

2.2 Galaxy formation in N-body simulations

2.2.1 Overmerging

One of the challenges for galaxy formation models is to define exactly what is meant by a galaxy 'halo'. Observationally, the term halo is applied to the outer part of the spheroidal mass component in a galaxy, including stars, gas and dark matter (baryonic or otherwise). In the CDM model, non-baryonic dark matter is the dominant form of matter in the universe; through gravitational collapse it forms the potential wells into which the baryons settle after recombination to form visible galaxies. In this context, a halo is to be thought of as a collapsed, virialised region of space. In practice, this is often taken to mean a region of space in which the mean overdensity is $\sim$ 200 times the critical density, the density contrast predicted by the spherical collapse model. However, dissipationless N-body simulations show that substructure survives in overdense regions, even in the dense cores of clusters (Ghigna et al. 2000). This is expected to be even more important when dissipation is taken into account, since the dissipational collapse of baryons compresses the dark matter, leading to dense cores that are difficult to destroy. The problem is also evident in the real world: rich galaxy clusters, which are expected to be fully relaxed, contain many tens, or even hundreds, of individual galaxies (which presumably correspond to dark matter subclumps) and do not resemble scaled-up galactic haloes, as expected (Moore et al. 1999).

The collapse and virialisation of dark matter haloes is a highly non-linear process and following it requires numerical techniques. N-body simulations use discrete particles to sample the dark matter distribution as it evolves under the influence of gravity alone. To follow the formation and merging of galaxy haloes, it is necessary to identify groups of particles representing such structures. The two most common techniques

for this purpose are the *spherical overdensity* approach (Warren et al. 1992; Lacey & Cole 1994) and *percolation algorithms* (also known as *friends-of-friends*; Davis et al. 1985). Percolation algorithms link particles pairwise to form groups; the criterion for linking two particles is that their separation is less than some threshold length, usually expressed as a fraction of the mean interparticle distance. Spherical overdensity, on the other hand, starts by identifying maxima in the density field and then places successively larger spheres around each maximum until the mean overdensity reaches some threshold $\delta \simeq 200$.

An alternative approach to following the merger histories of haloes, employed by semi-analytic models (SAMs), uses Monte–Carlo realisations of the extended Press–Schechter formalism (Bond et al. 1991; Bower 1991). The Press–Schechter theory is a linear theory and so does not explicitly include the collapse of dark matter haloes, which is a highly *non*-linear process. In the Bond et al. approach, the linear density field around each point in space is smoothed over successively larger spheres to find the largest sphere for which the mean overdensity $\delta > \delta_c$, the critical density expected for virialisation. By taking the largest spherical region, the problem of haloes embedded in larger overdensities (the 'cloud-in-cloud' problem) is avoided. The extended Press–Schechter formalism has found common use because it matches the spectrum of halo masses found in N-body simulations (e.g. Jenkins et al. 2001), which include non-linear effects implicitly. However, both the analytic technique and the halo finding algorithms described above suffer from 'overmerging' — a failure to identify substructure in overdense regions.

Consider the two situations shown in Fig. 2.1, which shows the sequence of events for two merging haloes. In the extended Press-Schechter formalism (bottom panel) all substructure is erased when the two haloes merge, and the galaxies contained within the original haloes are released into the new halo. Over time, their orbits will decay through dynamical friction and they will merge with the central galaxy. If, however, the original haloes survive as substructure within the new halo (top panel) then the galaxies within each subhalo can be expected to merge with one another on a much shorter timescale than the timescale to merge with the central galaxy in the former case. The result is a halo without a single dominant galaxy.

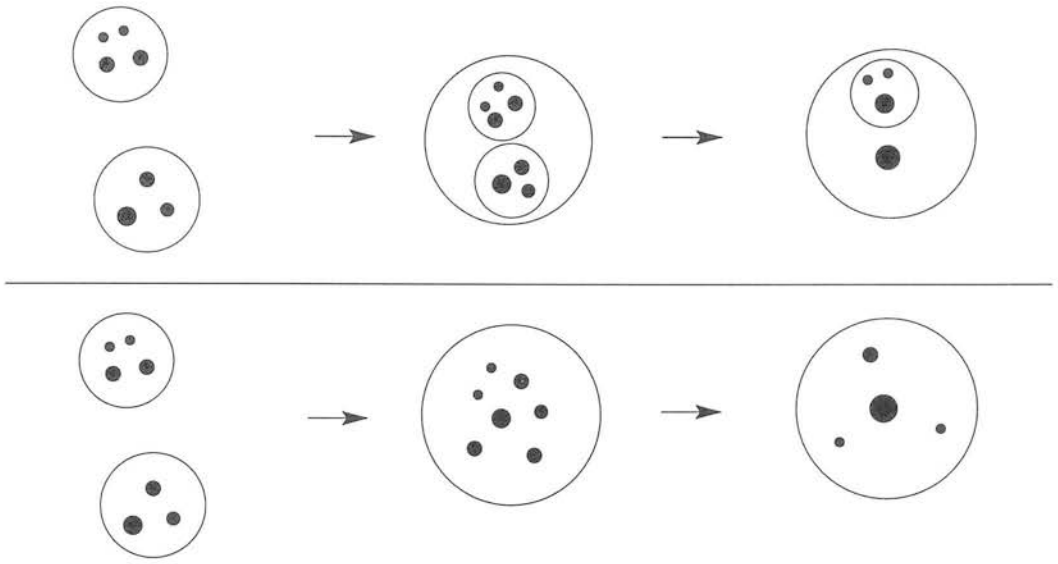

Figure 2.1: A sketch illustrating the problem of overmerging in galaxy formation models. Top panel: a typical sequence of galaxy merging within a hierarchy of subhaloes, as in the model presented in this thesis. Bottom panel: the predicted sequence of merging in a model based on the Press-Schechter formalism, in which all substructure is erased on formation of a new halo. Figure taken from van Kampen & Rimes (in preparation).

In this section I describe a method for incorporating the formation of galaxy haloes into an N-body simulation. Overdense groups of particles are identified using *local* density percolation; if a group satisfies a condition for virialisation it is replaced by a single softened particle (van Kampen 1995). This is intended to preserve substructure within overdense regions by preventing disruption of groups by 2-body heating and artificial tidal stripping. The former is a numerical effect due to the discrete nature of the N-body particles used to sample the matter distribution, while the latter is a real physical effect which is nevertheless overestimated in purely gravitational simulations, which have no dissipative matter component. Dissipation results in haloes with dense cores which are less easily disrupted, while the use of softened particles results in haloes that are artificially 'puffed up' and more easily stripped. By applying the galaxy formation algorithm many times during the evolution of the simulation, the formation and merging of halo particles can be followed, resulting in a merger history, or tree, for each present-day halo, which takes account of the existence of substructure explicitly.

In the following discussion, I will use the term *halo* to describe a region of space

with overdensity $\delta \simeq 200$, such as would be identified as a halo by percolation with a fixed linking length. This definition is the one most commonly adopted in galaxy formation models and corresponds to the Press-Schechter definition of a halo. I will use the term *subhalo* to mean a fully virialised overdensity, capable of cooling gas and forming a visible galaxy. A subhalo may reside within a larger (as yet unvirialised) halo, or it may be a halo in its own right (in which case I will refer to it as either a halo or a subhalo, depending on the context).

2.2.2 Defining subhaloes

LOCAL DENSITY PERCOLATION

Percolation algorithms are commonly used for identifying galaxy haloes in N-body simulations. Overdense regions of space are picked out by linking up groups of dark matter particles that have pairwise separations

$$r_{ij} \leqslant b\langle n\rangle^{-1/3}, \tag{2.1}$$

where $\langle n\rangle$ is the mean particle density over the whole simulation volume. In traditional friends-of-friends the linking parameter, b, is fixed and is chosen to identify structures with average overdensities of $\delta = \delta n/\langle n\rangle \simeq 100 - 200$, the density contrast expected for virialised structures in the spherical collapse model. Fig. 2.2 shows a 2-d example, in which a percolation algorithm with $b = 0.2$ (a common choice in cosmological simulations) is applied to an artificial overdensity, generated by randomly placing particles according to a Gaussian distribution superimposed on a uniform background, with the relative levels adjusted so that an overdensity of $200^{2/3} \simeq 30$ (because there is one less dimension than the usual 3-d case) is expected within the dotted circle. The algorithm picks out most of the particles within the overdense region (as well as a few outside); the final group contains 109 particles, slightly more than the 105 expected from the input distribution. However, one failure of this method is that it fails to separate substructure in regions of high density (in a sense this is the opposite of the cloud-in-cloud problem, which is only a problem if one assumes that all substructure is erased), while linking together oddly-shaped groups in low-density environments by percolating across bridges and filaments. This was first realised by Bertschinger & Gelb

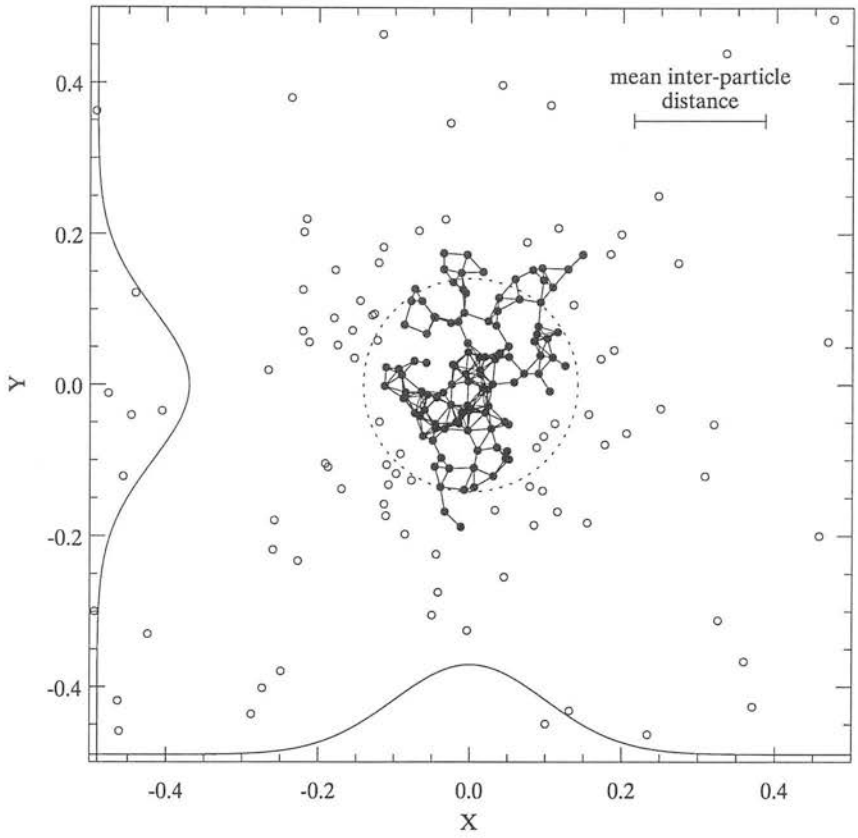

Figure 2.2: An example of a percolation algorithm at work. Particles (filled and open circles) have been distributed randomly according to a Gaussian distribution with $\sigma_X = \sigma_Y = 0.1$, superposed on a uniform background in such a way as to give an overdensity of ~ 30 above the background within a circle of radius σ (marked by a dotted circle). Filled circles mark particles identified as belonging to the largest group (and the only one with > 10 members) by the percolation algorithm and solid lines mark the actual links made. A fixed linking parameter $b = 0.2$ was used.

(1991); their DENMAX algorithm (and, later, its close relative SKID; Governato et al. 1997) was one of the first attempts to deal with this problem, by temporarily evolving the particle distribution towards local density maxima, to separate out haloes in high density regions.

In local density percolation (which could be called adaptive friends-of-friends) the linking length is modified by the local density, so that groups of particles are linked if they are separated by some fraction of what would be expected locally. For a Poissonian

distribution, the mean nearest-neighbour distance is

$$r_{\rm nn}^{\rm p} = (4\pi\langle n\rangle/3)^{-1/3}. \tag{2.2}$$

Note the factor of $(4\pi/3)^{-1/3}$ difference between this definition and equation (2.1). The local linking length for equal-mass particles is given by

$$R_0(\mathbf{x}_i, p, s) = p\left[\frac{4\pi n_{\rm G}(\mathbf{x}_i, s)}{3}\right]^{-1/3} \tag{2.3}$$

$$= pr_{\rm nn}^{\rm p}\left[\frac{n_{\rm G}(\mathbf{x}_i, s)}{\langle n\rangle}\right]^{-1/3} \tag{2.4}$$

where $n_{\rm G}$ is the local density smoothed with a Gaussian of width $R_{\rm s} = sr_{\rm nn}^{\rm p}$:

$$n_{\rm G}(\mathbf{x}_i, s) = (2\pi R_{\rm s}^2)^{-3/2}\int n(\mathbf{x}')\exp\left(-|\mathbf{x}_i - \mathbf{x}'|^2/2R_{\rm s}^2\right)\,{\rm d}^3\mathbf{x}'. \tag{2.5}$$

A further complication arises when groups of particles are replaced by single, more massive particles. To account for the presence of different mass particles in the simulation, a particle with mass m_i, an integer multiple of the initial mass m_0, is placed a distance $\sqrt{m_i/m_0}$ further away, where it exerts the same gravitational force as a particle of mass m_0. So the percolation length for particle i is

$$R(\mathbf{x}_i, p, s) = pr_{\rm nn}^{\rm p}\sqrt{\frac{m_i}{m_0}}\left(\frac{n_{\rm G}(\mathbf{x}_i, s)}{\langle n\rangle}\right)^{-1/3}. \tag{2.6}$$

With local density percolation, each particle has its own percolation length, so for each pair of particles the mean of the two is taken. To prevent large groups from being linked together in low density regions, due to excessively large linking lengths, a maximum of $0.32r_{\rm nn}^{\rm p}$ is adopted:

$$R_{ij} = \min\left[\frac{R(\mathbf{x}_i, p, s) + R(\mathbf{x}_j, p, s)}{2}, 0.32r_{\rm nn}^{\rm p}\right]. \tag{2.7}$$

This upper limit was chosen to ensure that the field population is the same as would be identified by percolation with a fixed linking length, enabling a fair comparison between the two techniques. This is the equivalent to requiring that local density percolation find substructure within haloes but not identify new haloes. The two free parameters of the model are thus s, the smoothing length, and p, the linking parameter.

VIRIAL TEST

We only want to replace groups by single particles if they represent virialised structures. Spitzer (1969) showed that many equilibrium systems satisfy a simple relation between their half-mass radii, $R_{\rm half}$, and velocity dispersions,$\langle v^2 \rangle^{1/2}$:

$$R_{\rm half} \simeq 0.4 \frac{GM}{\langle v^2 \rangle}. \tag{2.8}$$

This is used to define a *virial estimator*,

$$V = 2.5 \frac{R_{\rm half} \langle v^2 \rangle}{GM}, \tag{2.9}$$

for each group of particles identified by the percolation algorithm. If V is in the range $1 - \delta V < V < 1 + \delta V$ then the group is considered to be virialised and is replaced by a single softened particle. The virial tolerance, δV, is the third free parameter of the model. The choice of softening length for the new halo particle determines its potential energy hence its internal kinetic energy, since the system is in virial equilibrium. The softening length, ϵ, is chosen to match the half-mass radius of the original group, so as to conserve energy. For Plummer softening,

$$R_{\rm half} \simeq 1.3\epsilon \tag{2.10}$$

(see appendix A). Here, $R_{\rm half}$ is taken to be the mass-weighted mean radius, since this is a less noisy estimator than the median. Halo particles, once formed, may themselves be incorporated into groups at later times. In this way, the merging of dark matter haloes is taken into account automatically.

A minimum of 10 particles is set on group membership. To prevent numerical problems caused by excessively large single particles, an upper limit of $2.6 \times 10^{11} h_0^{-1}$ $\mathrm{M}_\odot$, or ~ 200 of the original particles, is placed on the mass of a halo particle. Groups more massive than this limit are not replaced by single particles. This is not likely to have a major effect on the stability of such groups, as groups this massive are not easily disrupted.

Table 2.2: Parameters of the group finders used to define haloes and subhaloes.

	Group finder	b	$p^\star$	$s^\star$	$\delta V^\star$
haloes	fixed	0.2			
subhaloes	adaptive+virial test		1.0	0.5	0.5

2.2.3 The subhalo population

SIMULATIONS

The simulation ensemble comprises five identical cubic N-body boxes, each with volume $(32h_0^{-1}$ Mpc$)^3$. The initial conditions for each box, realisations of an unconstrained, Gaussian random field, were generated using the algorithm of Hoffman & Ribak (1991), as implemented by van de Weygaert & Bertschinger (1996). Initially, each box contained 128^3 dark matter particles with masses of $1.3 \times 10^9 h_0^{-1}$ $M_\odot$. The boxes were evolved using a standard tree code (Barnes & Hut 1986), supplemented by the galaxy formation algorithm outlined above. Periodic boundary conditions were made possible by implimentation of the Ewald (1921) summation method, as described in Hernquist, Bouchet & Suto (1991).

The galaxy formation algorithm was applied ~ 40 times, at 'snapshots' uniformly spaced in $a = 1/(1+z)$. The parameters used to identify virialised groups for the galaxy formation algorithm were $p = 0.5$, $s = 1.0$ and $\delta V = 0.25$. This choice of parameters is deliberately conservative; the idea is to prevent disruption of tightly-bound groups by purely numerical effects without artificially preserving substructure which *should* be destroyed. By choosing a small linking length and large smoothing length, and by applying the virial test strictly, we ensure that only the dense cores of haloes, where dissipation is important, are replaced by single particles, while some stripping of the outer, less tightly bound parts of haloes is allowed.

In Figs. 2.3 and 2.4 I compare the groups identified using different halo finding algorithms. The top box of Fig. 2.3 shows the dark matter distribution in one of the five simulation boxes at the final snapshot. The bottom box shows groups of particles identified by percolation using a fixed linking length of $b = 0.2$. Each group is

plotted as a solid sphere with a diameter that scales with $M^{1/3}$, which is approximately proportional to linear size; the overall scaling of the group sizes is arbitrary and was chosen to give roughly the right size for the largest groups. These groups are the 'haloes', as defined by the spherical collapse model. The large-scale structure in the dark matter distribution is clearly mirrored in the halo population, with the largest haloes being strung out along the filamentary structures seen in the top box.

The top box of Fig. 2.4 shows all groups identified by local density percolation. The linking parameters used in this case are listed in the second row of table 2.2. These choices are less conservative than those used in the galaxy formation algorithm; this is designed to identify the whole of each subhalo, rather than just the dense core. The notation '*' is used to distinguish the parameters used here from the equivalent parameters used in the galaxy formation algorithm. In the bottom box of Fig. 2.4 I show only groups that satisfy the virial criterion. Where a group is not virialised, it is replaced by its last virialised progenitors. The cores of the most massive haloes are still visible in the subhalo distribution, but the central virialised region contains only a fraction of the mass of the whole halo. The most massive haloes will contain rich galaxy clusters and, in this context, the virialised core corresponds to the region from which gas can cool onto the central cD galaxy. Substructure which still survives in the outer parts of the fixed linking length groups is identified as separate groups by local density percolation.

Note that in all of the three cases outlined above, the simulation was run using the full galaxy formation recipe, with parameters as described in section 2.2.3; the only difference is in the method used to identify the final (sub)halo population.

Subhalo mass function

The consequences of using different group finders is illustrated by Fig. 2.5, which shows the differential mass functions of haloes and subhaloes at the present day. The predictions of Press & Schechter (1974, light, dashed curve) and Sheth & Tormen (2002, dash-dotted curve) are also plotted, along with the N-body mass function of Jenkins et al. (2001, light, solid line). All of the model curves (heavy lines) have been smoothed with a Gaussian of width 0.1 in $\log_{10} M$, as in Jenkins et al. The dotted line is for

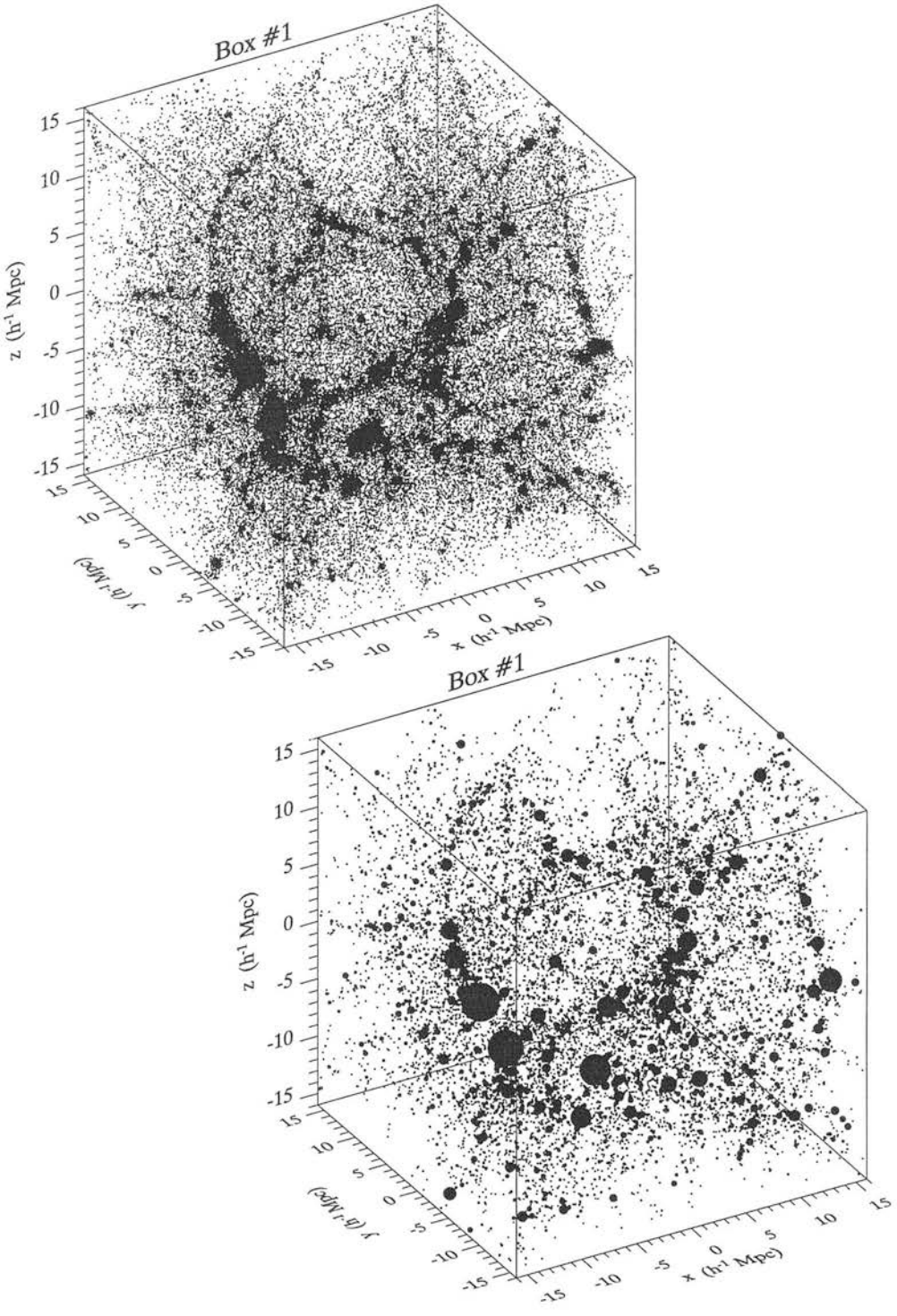

Figure 2.3: Top panel: the dark matter distribution in one of the five simulations at $z = 0$. Only every tenth dark matter particle has been plotted, for clarity. Bottom panel: the distribution of groups identified by percolation with a fixed linking length (the haloes). Each dot represents a group of particles (or a single halo particle). The diameters of the dots are scaled by $M^{1/3}$ (which is approximately proportional to linear size).

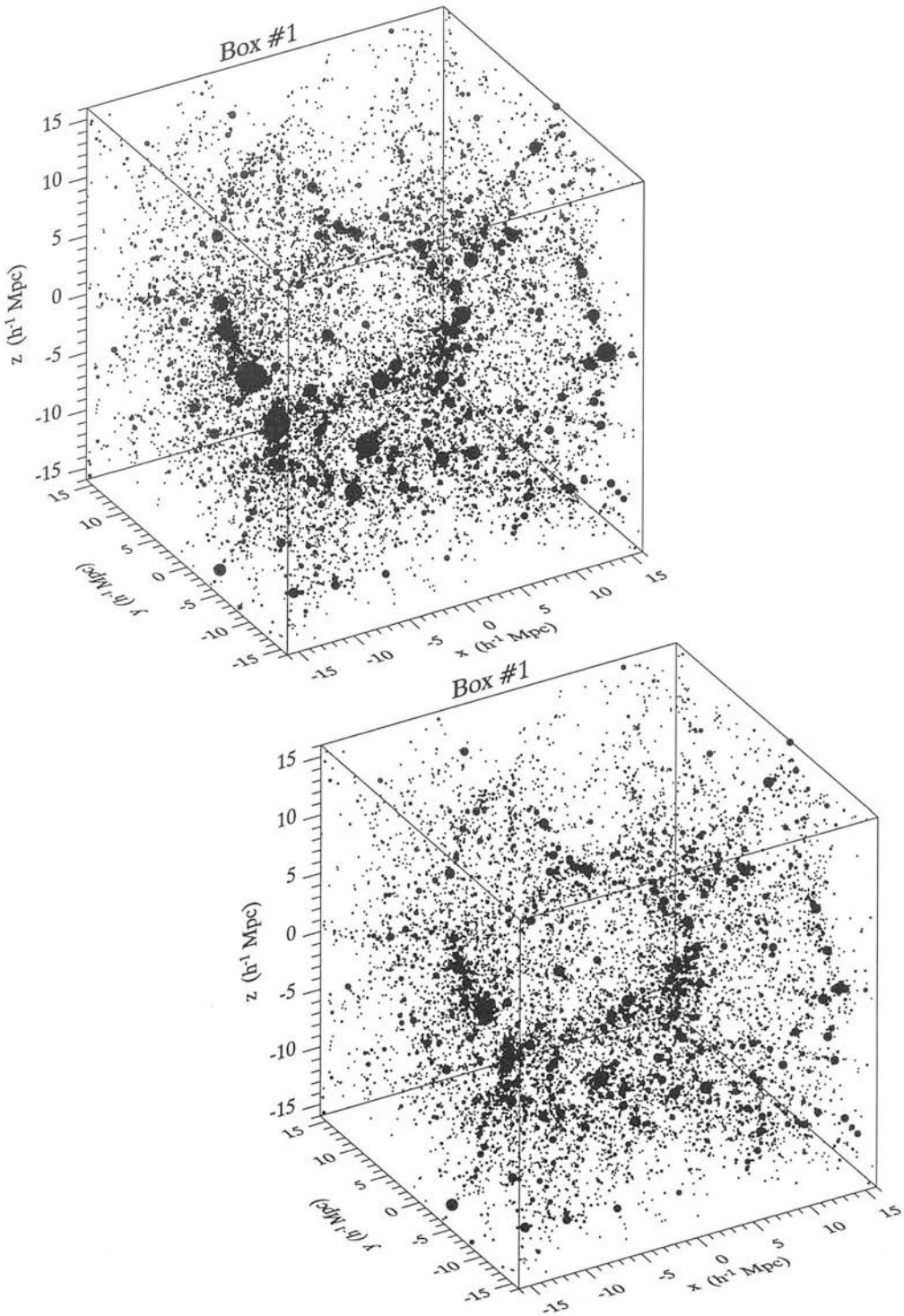

Figure 2.4: Top panel: as the bottom panel of Fig. 2.3, but this time for groups identified by local density percolation, with parameters as listed in table 2.2. Bottom panel: groups identified by local density percolation and satisfying the virial criterion of equation 2.8 (the subhaloes). Groups that do not satisfy this criterion are split up into their progenitors.

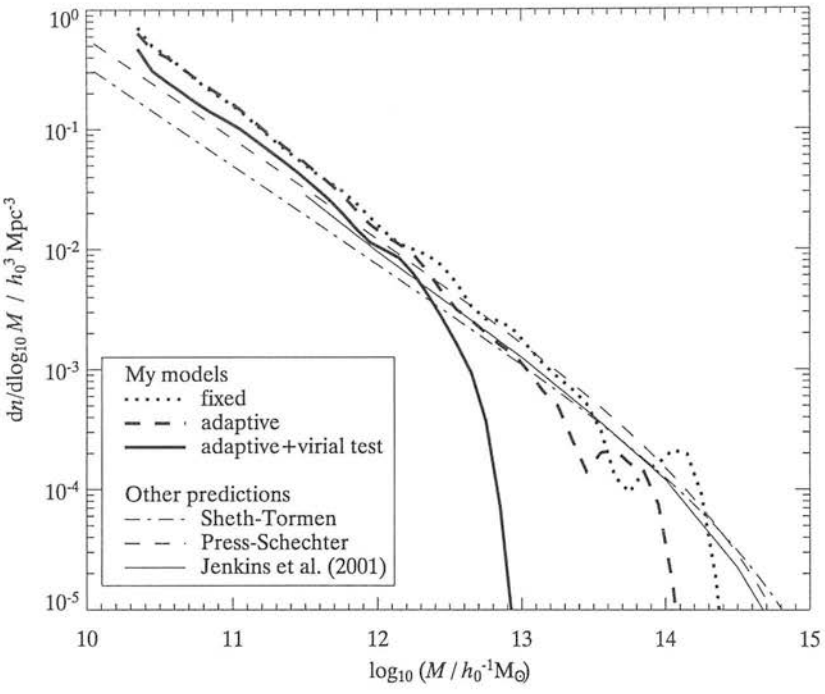

Figure 2.5: The differential halo/subhalo mass function for one of the five simulations. The heavy lines are for groups identified in the simulations presented in this work. The dotted line is for groups found using percolation with a fixed linking length (the haloes), while the dashed line is for groups identified using local density percolation, as described above. The solid line is for the subset of the latter that satisfy the virial test (the subhaloes). The light solid line is the ΛCDM mass function of Jenkins et al. (2001) and the lighter dashed and dash-dotted lines are halo mass functions from analytic theory, using the methods of Press & Schechter (1974) and Sheth & Tormen (2002) respectively.

haloes identified using percolation with a fixed linking length, $b = 0.2$. It compares well with the mass function of Jenkins et al. in the overlapping mass range (the cut-off at high masses is a result of the smaller box size used in this work: $32^3 h_0^{-3}$ Mpc^3, as compared with their $3000^3 h_0^{-3}$ Mpc^3). Comparing the mass function obtained using local density percolation (heavy dashed line) with that obtained using a fixed linking length, it can be seen that the most massive haloes in the simulation, identified as single objects by traditional friends-of-friends, are split up into smaller subhaloes, resulting in fewer massive haloes. However, many of these subhaloes do not satisfy the virial test (equation 2.9). In these cases, the progenitors of these subhaloes are retained as

distinct entities, further reducing the number of massive galaxy haloes and resulting in the solid line in figure 2.5. This is correct since the largest collapsed objects in the universe, galaxy clusters, are typically made up of many tens, or even hundreds, of individual galaxies and do not neccessarily contain a single dominant galaxy.

Virial properties

For each group of particles in the N-body simulation we can measure the following quantities: its total mass, M; its half-mass (or, in practice, the median) radius, $r_{\rm half}$; and the r.m.s. velocity dispersion, $\langle v^2 \rangle^{1/2}$. The spherical collapse model, on the other hand, makes predictions for the virial radius, r_{200} (equation 1.41), and the virial velocity, V_{200} (equation 1.39). For the simple case of a singular isothermal sphere, these quantities can be related to the half-mass radius and velocity dispersion as follows. The virial radius of a truncated isothermal sphere is, by definition, $r_{\rm vir} = 2r_{\rm half}$ (since $M(r) \propto r$). The virial velocity can be related to the velocity dispersion through the internal energy:

$$E = -\frac{GM^2}{2r_{\rm vir}} = -\frac{MV_{\rm c}^2}{2}, \tag{2.11}$$

which implies

$$V_{\rm c} = \langle v^2 \rangle^{1/2} \tag{2.12}$$

(it interesting to note that this is the result that would be expected if all of the particles were on circular orbits with speed $V_{\rm c}$, since a particle on a stable circular orbit in an inverse-square law potential trivially satisfies the virial theorem). In Figs. 2.6 and 2.7 I compare the properties measured in the simulations with the predictions of the spherical collapse model. The dashed lines in each panel correspond to the expected relationships. Fig. 2.6 makes the comparison for all groups identified by percolation with a fixed linking length (haloes), while Fig. 2.7 is for all virialised groups identified by local density percolation (subhaloes). In addition to the properties defined above, I also plot two derived quantities: $\tau_{\rm dyn} = 2r_{\rm half}/\langle v^2 \rangle^{1/2}$, a measure of the dynamical time of the system, and δ, the spherical overdensity. Again, these are compared to the equivalent quantities predicted by the spherical collapse model: $\tau_{\rm orbit} = \pi r_{200}/V_{200}$ and δ_{200} (equation 1.35).

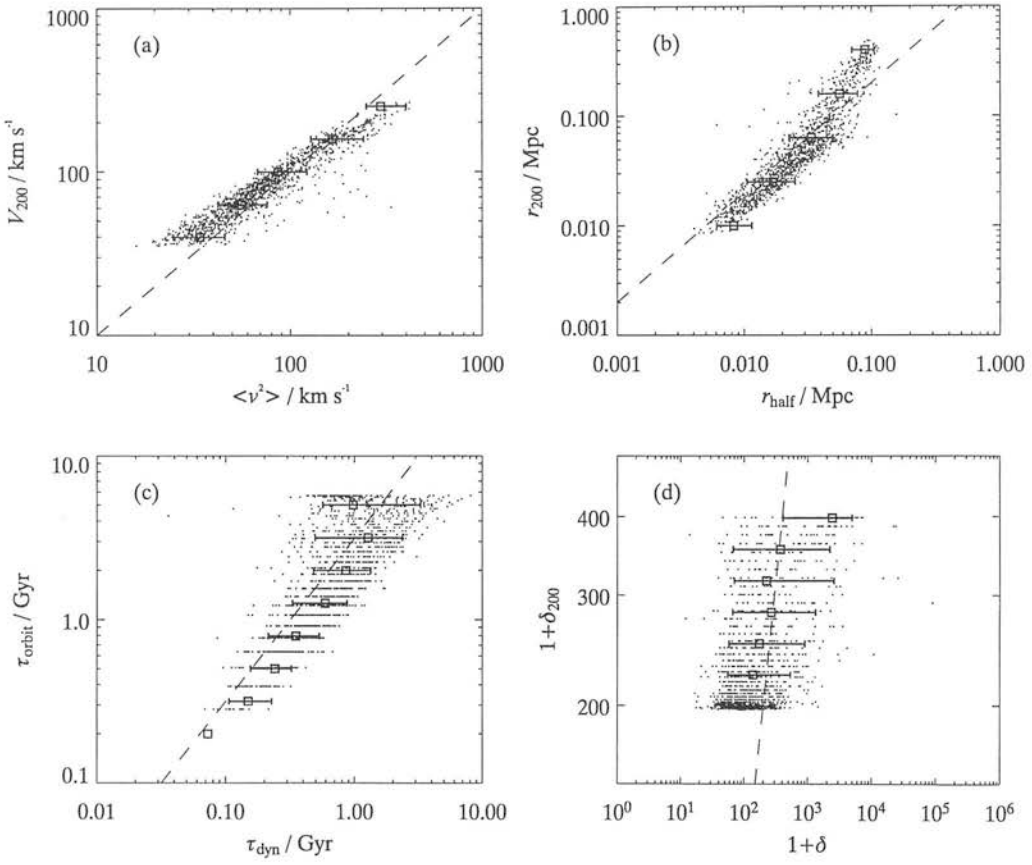

Figure 2.6: A comparison of the properties of friends-of-friends haloes in the N-body simulation with the predictions of the spherical collapse model: (a) the velocity dispersion, (b) the half-mass radius, (c) a measure of the dynamical time and (d) the spherical overdensity. Only every tenth halo is plotted, for clarity. In each panel the squares with error bars denote the median and 10 and 90% quartiles of the distribution of simulated halo properties and the dashed line marks the relation expected for the simple case of a singular isothermal sphere.

Both haloes and subhaloes give a good match to the properties predicted by the spherical collapse model *on average*. However, even for the haloes there is considerable scatter about the expected properties — this intrinsic scatter in the properties of galaxy haloes is ignored in analytical and semi-analytic galaxy formation models. The only notable difference between the properties of the two populations is the absence of any high velocity dispersion subhaloes and the tendency of the most massive subhaloes to have lower velocity dispersions than would be expected from the spherical collapse model. The trend of increasing overdensity with time for a Λ-dominated universe (such

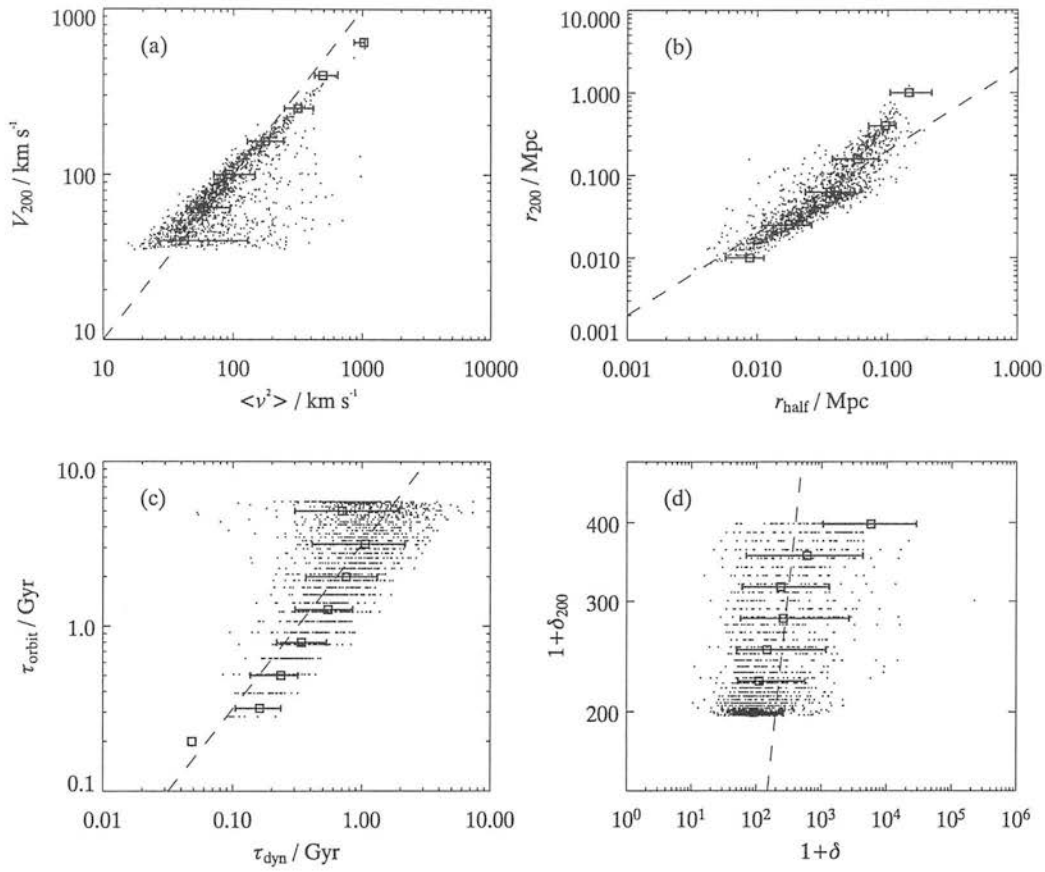

Figure 2.7: As Fig. 2.6 for virialised subhaloes.

that the expected overdensity of haloes at the present day is nearer 400 than 200) is seen in the simulations, albeit with a large scatter. Note that the discreteness seen in the bottom two panels is the result of the small number of snapshots.

As an aside, we can use equations 2.11 and 2.12 to write the virial theorem for the singular isothermal sphere as

$$r_{\rm half} \simeq 0.5 \frac{GM}{\langle v^2 \rangle}. \tag{2.13}$$

This is clearly inconsistent with equation (2.8), which is valid for more general density profiles (as shown by Spitzer 1969), so the virial estimator used in the galaxy formation algorithm does not strictly apply in this case. However, the difference between the two expressions is within the scatter of $\pm 50\%$ allowed by the virial test, so the effect on the average properties of haloes is small.

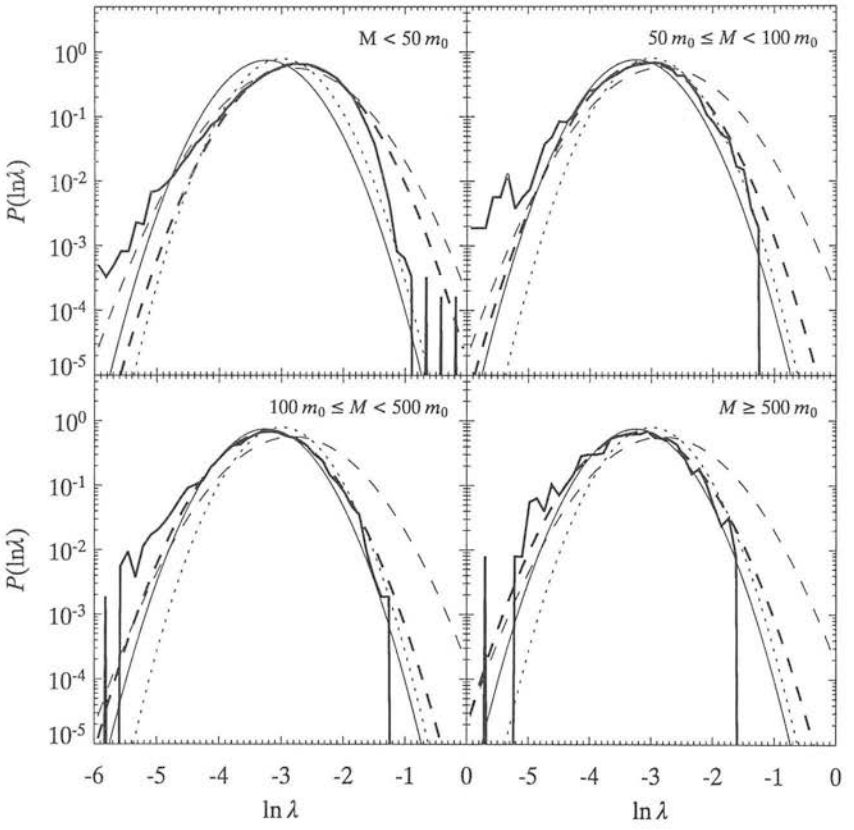

Figure 2.8: The distributions of subhalo spin parameters measured in the N-body simulations, for four mass ranges. The masses are expressed in terms of the N-body particle mass, $m_0 = 1.3 \times 10^{10} h_0^{-1}$ $M_\odot$. The heavy, solid histogram in each panel is the distribution of spin parameters measured in the simulation and the heavy, dashed curves are log-normal fits to these histograms. The lighter curves are spin distributions measured by other groups (Cole & Lacey 1996, solid curve; Mo, Mao & White 1998, dotted curve; Dalcanton, Spergel & Summers 1997, dashed curve; the last two are both based on the results of Warren et al. 1992).

Spin parameters

The angular momentum of a system can be conveniently parameterised by the dimensionless spin parameter

$$\lambda \equiv \frac{J|E|^{1/2}}{GM^{5/2}}. \tag{2.14}$$

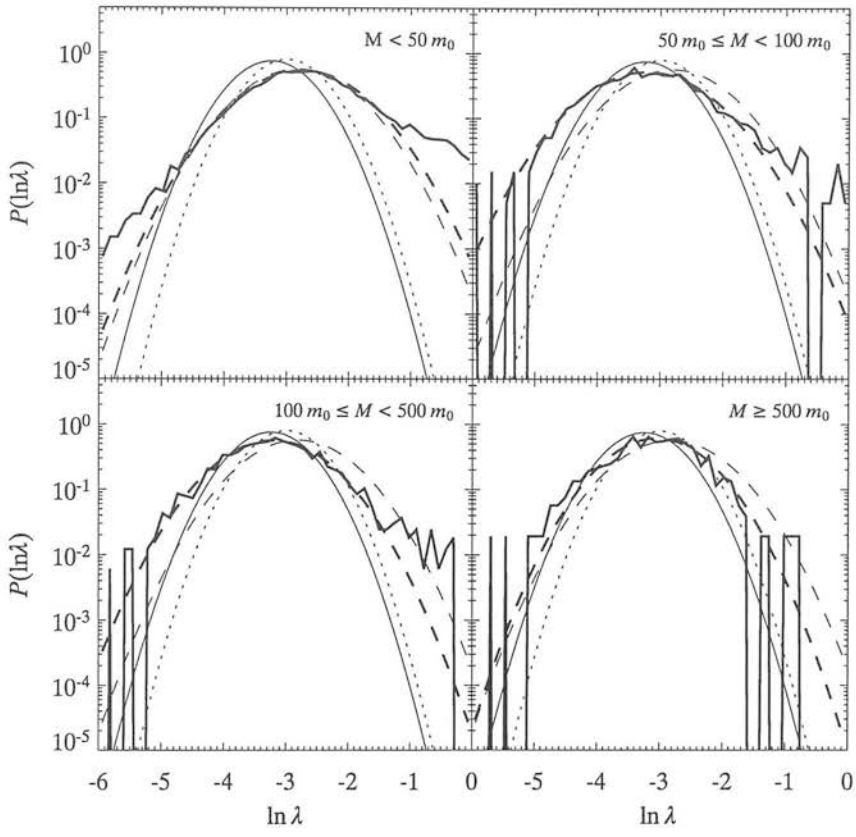

Figure 2.9: As Fig. 2.8 for friends-of-friends haloes.

where M is the total mass of the system, J is the angular momentum and E is the binding energy. For groups of particles, these are given by

$$M = \sum_i m_i, \qquad \mathbf{J} = \sum_i m_i \mathbf{r}_i \times \mathbf{v}_i, \qquad E = -\frac{1}{2} \sum_i m_i |\mathbf{v}_i|^2. \tag{2.15}$$

In Fig. 2.8 I plot the distribution of subhalo spin parameters measured in the simulations, for four different mass ranges. The masses are expressed in terms of the equivalent number of original N-body particles, to give some idea of the types of groups in each bin. The actual number of particles in a group may be less because the group may contain one or more halo particles. The distributions are well-fitted by lognormal functions (heavy, dashed curves) with median in the range $\lambda_{\rm med} = 0.040$–$0.043$, except for the least massive haloes, which have $\lambda_{\rm med} = 0.063$. The latter result is almost

certainly due to numerical effects: the lowest mass bin is dominated, numerically, by galaxies with the minimum number of particles (10) and these are severely affected by statistical fluctuations and by 'interlopers', particles that are not bound to the group, but by chance are included at a given time step[1]. Fig. 2.9 shows the spin distributions for friends-of-friends haloes. The median spin parameters for haloes and subhaloes agree well and both agree with what is found for haloes in high-resolution N-body simulations (e.g. Cole & Lacey 1996; Mo, Mao & White 1998; Dalcanton, Spergel & Summers 1997; Lemson & Kauffmann 1999). Fits to the spin distributions in Figs. 2.8 and 2.9 yield widths in the range $\sigma_{\ln\lambda} = 0.59$–$0.61$ for the subhaloes, with friends-of-friends haloes having somewhat wider distributions: $\sigma_{\ln\lambda} = 0.65$–$0.77$. Note that the distributions in Fig. 2.9 in this plot are more noisy, because they are based on only two simulation boxes (data for fixed linking length haloes was only available for these two). This may contribute to the increased widths derived from the functional fits; however, restricting the analysis of the subhalo spin distributions to two boxes does not change the results significantly, except in the most massive bin, where there are too few subhaloes to make a reliable fit.

2.2.4 Subhalo merger trees

REMOVING DISCRETENESS EFFECTS

Because of the small number (~ 40) of snapshots, the time resolution of the merger trees produced by the above method is quite coarse (0.35 Gyr at $z = 0$). This is the one major disadvantage compared to analytic merger trees in which the time resolution is determined by the merger rate and can be different for different haloes. Discreteness effects are particularly important at recent times, when the exact ages of stellar populations become important and the time between snapshots is much larger than the star formation timescale. To minimise these effects, I assign each subhalo a random formation redshift, $z_{\rm form}$, between the snapshot, i, when it was first identified and the previous snapshot $(i-1)$, when it was not identified as a single, virialised structure. The snapshots are spaced uniformly in $a = 1/(1+z)$, so I pick the formation redshift

[1]The spins of large haloes can also be affected by numerical problems, mainly resulting from periodicity in the large scale structure caused by the boundary conditions (Warren et al. 1992).

from a top-hat distribution in a:

$$P(a_{\rm form}) = \begin{cases} 0 & a_{\rm form} \leqslant a_{i-1} \\ \frac{1}{a_i - a_{i-1}} & a_{i-1} < a_{\rm form} \leqslant a_i \\ 0 & a_{\rm form} > a_i \end{cases} , \tag{2.16}$$

where a_i is the expansion factor of the universe at snapshot i. The formation redshift of the halo is then $z_{\rm form} = (1/a_{\rm form}) - 1$.

EXAMPLE MERGER TREES

In Figs. 2.10 and 2.11 I plot four examples of subhalo merger trees. Each subhalo in the trees is represented by a circle, with diameter scaling as $M^{1/3}$, at the redshift when it formed. The most massive progenitor of each present day subhalo is marked by a star. Solid lines show the path by which each of the original progenitors is incorporated into the final subhalo. Subhaloes form in two ways: either by the merging of several progenitor subhaloes or, in the case where there is only one progenitor, by smooth accretion of mass onto an existing subhalo. Thus, each circle represents one of these types of event. Below each tree, I plot the fraction of the final mass in resolved, virialised progenitors (solid line) and in the most massive progenitor (dotted line) as a function of redshift.

I have chosen the four examples to illustrate the wide variation in the merger histories of present-day galaxies. The two subhaloes shown in Fig. 2.10 have comparable final masses, but in case (b) the final mass was built up slowly, mostly through smooth accretion, while most of the final mass in (a) (which is, in fact, the most massive subhalo in the simulation) came together in a major merger event (with significant accretion) at $z \simeq 0.5$. Fig. 2.11 shows the opposite end of the mass spectrum: in (a) I plot the merger history of a low-mass subhalo that formed in a single merger at $z \simeq 3.5$ and has grown very little since, while in (b) I show a subhalo that formed very early in the simulation and has since grown only by smooth accretion.

2.3 Basic ingredients for galaxy formation

For this section only, I drop the prefix from subhalo and simply use the term *halo* to refer to the elements of the merger trees. This is mainly done for convenience and also

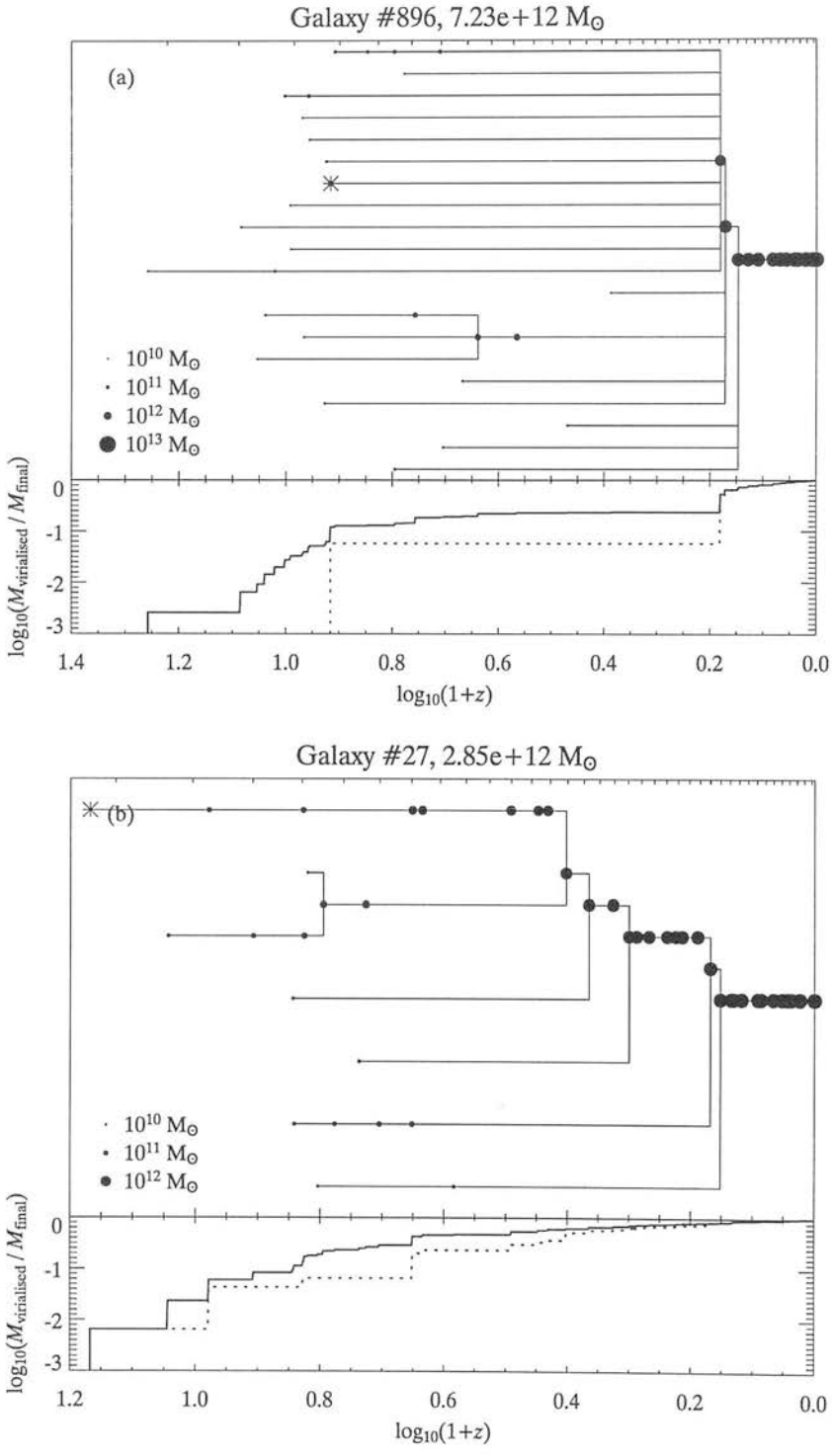

Figure 2.10: Examples of subhalo merger trees. The top panel in each plot shows the distribution of progenitors of the present day subhalo as a function of redshift. The sizes of the circles scale as $M^{1/3}$ and solid lines indicated the order of merging. In the bottom panels, the fraction of the final mass is plotted as a function of redshift, for all resolved, virialised progenitors (solid line) and the most massive progenitor (dotted line), which is marked with a star.

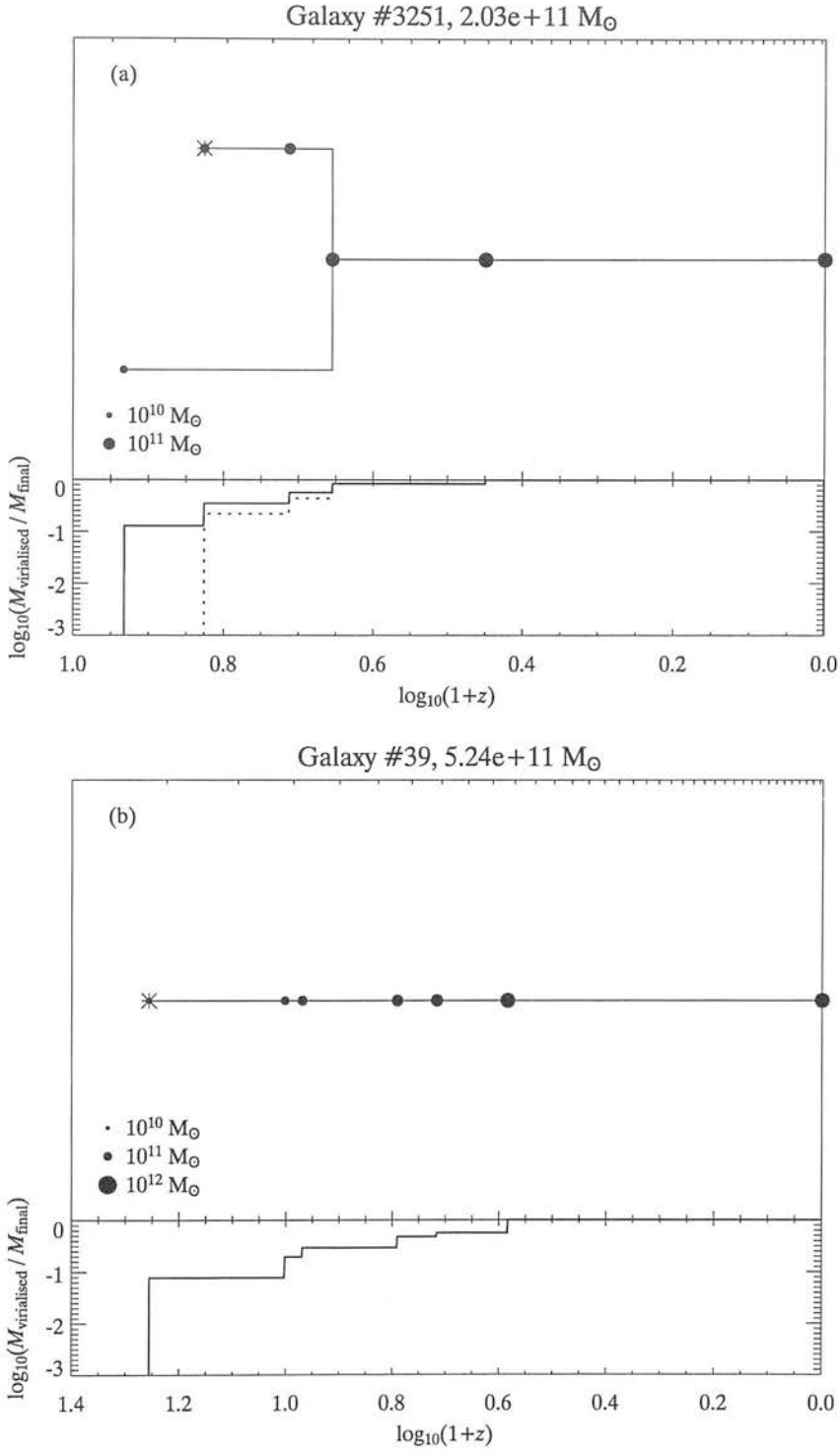

Figure 2.11: Further examples of subhalo merger trees. For a detailed description see Fig. 2.10.

because the model presented here can be applied equally to merger trees consisting of haloes or subhaloes.

2.3.1 Galaxy merging

Each new halo initially forms a single galaxy at its centre; when haloes merge, the outer regions of the progenitor haloes are disrupted but the baryonic cores survive for some time as distinct entities. In the following discussion, I use the term *galaxy* to refer to the gas, stars and any dark matter associated with an individual baryonic condensation. In addition to the dominant central galaxy, a halo may contain several satellite galaxies. These satellite galaxies orbit the centre of the new halo, losing orbital energy through dynamical friction until they merge with the central galaxy in the halo, or until the halo in which they are orbiting is incorporated into a more massive halo.

To decide which galaxies merge with the central galaxy, a merger timescale, $\tau_{\rm mrg}$, is calculated for each satellite using dynamical friction arguments, as described below. Those satellites for which $\tau_{\rm mrg}$ is less than the time available for merging are merged with the central galaxy at the time of the halo merger event. The stars in these galaxies are placed in the bulge of the central galaxy and their reservoirs of hot and cold gas are merged with the respective reservoirs of the central galaxy. I assume, for the moment, that the disc of the central galaxy is unperturbed by the merger, which should be the case if the satellite is much smaller than the central galaxy (Walker, Mihos & Hernquist 1996), but not if the masses of the two galaxies are comparable (e.g. Barnes 1992). I also ignore, for now, the possibility that significant quantities of gas will be driven into the centre of the surviving galaxy by the merger, fuelling a nuclear starburst.

If a satellite does not merge before the end of the halo lifetime then it keeps the properties (circular velocity, angular momentum, etc.) that it had in the previous halo. Thus, if a galaxy survives as a satellite through a succession of different halo mergers, its original properties are preserved until such time as it does merge, or until the present day.

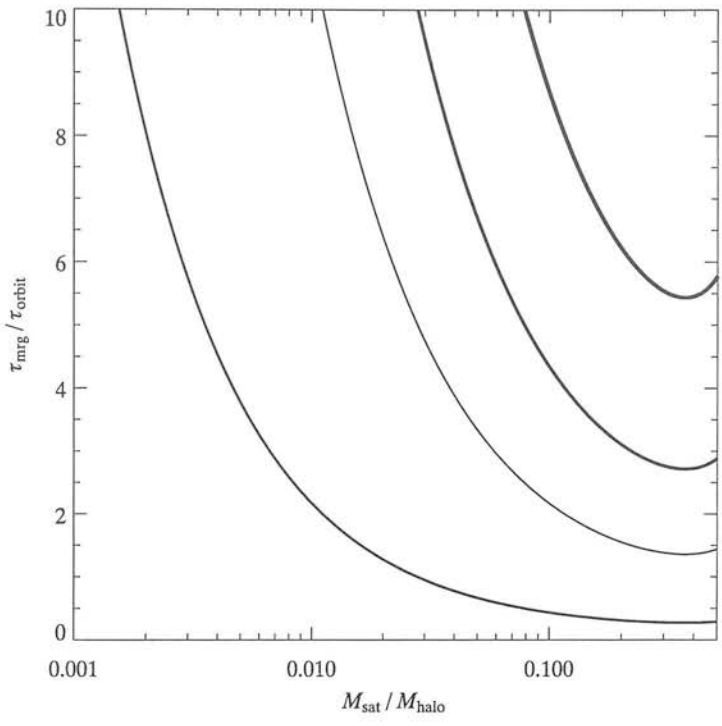

Figure 2.12: Dynamical friction timescale as a function of the ratio of the masses of the satellite galaxy and its parent halo. Curves are plotted for (top to bottom) $f_{\rm mrg} = 2$, 1, 0.5 and 0.1.

Dynamical friction

The time needed for a point-mass satellite, starting on a circular orbit at the virial radius, to lose all of its orbital energy is given by the Chandrasekhar (1943) formula and can be shown to be

$$\tau_{\rm mrg} = \frac{1.17 r_{\rm vir}^2 V_{\rm c}}{G M_{\rm sat} \ln(M_{\rm halo}/M_{\rm sat})}, \tag{2.17}$$

(Binney & Tremaine 1987, equation 7-26), where $M_{\rm halo}$ is the mass of the parent halo and $M_{\rm sat}$ the mass of the satellite. Since this expression relies on approximating the satellite as a point mass, it is only strictly valid for small ($M_{\rm sat} \ll M_{\rm halo}$), compact satellites. I take the mass of the satellite to be that of the halo in which the satellite previously resided; implicit in this is the assumption that the dark halo of the satellite is not stripped. Equation (2.17) can be rewritten as

$$\tau_{\rm mrg} = \frac{0.3722 M_{\rm halo}/M_{\rm sat}}{\ln(M_{\rm halo}/M_{\rm sat})} \tau_{\rm orbit}, \tag{2.18}$$

where $\tau_{\rm orbit}$ is defined to be half the orbital time at the virial radius (or approximately the orbital time at the half-mass radius for most density profiles):

$$\tau_{\rm orbit} = \frac{\pi r_{\rm vir}}{V_{\rm c}}. \tag{2.19}$$

Equation (2.18) is an upper limit to the merging timescale: a galaxy starting from a random orbit will typically fall to the centre of the halo in a shorter time (a factor of ~ 0.7 on average; CLBF).

The procedure for obtaining galaxy merger histories from those of their parent haloes is as follows. For each galaxy I calculate a merger timescale using the general expression

$$\tau_{\rm mrg} = f_{\rm mrg}\tau_{\rm dyn}\frac{M_{\rm halo}/M_{\rm sat}}{\ln(M_{\rm halo}/M_{\rm sat})}, \tag{2.20}$$

where I leave $f_{\rm mrg}$ as a free parameter. This is plotted in Fig. 2.12 for different values of $f_{\rm mrg}$. The curves are cut off at $M_{\rm sat}/M_{\rm halo} = 0.5$; there is an unphysical singularity at $M_{\rm sat} = M_{\rm halo}$, resulting from a breakdown in the approximations used in deriving the equation, specifically the assumption of a point mass orbiting in a fixed potential. By definition, no satellite can have $M_{\rm sat} > 0.5M_{\rm halo}$, since this would make it the most massive, and hence the central, galaxy in the halo. In practice, almost all satellites have $M_{\rm sat} \ll M_{\rm halo}$.

A first estimate of the time available for merging is the lifetime, $t_{\rm life}$, of the new halo. However, this relies on the assumption that all galaxies enter the halo on random orbits (here, I take random to mean simply that equation 2.20 is a good estimate of the time taken for an orbit to decay). The validity of this assumption depends on the nature of the merger event in which the new halo formed. Consider two extreme cases: if a massive halo merges with a much smaller one to form a new halo then the orbits of satellites in the larger of the two progenitors can be expected to be relatively unperturbed by the event (in fact, it could be argued that this does not constitute a new halo at all, but simply a continuation of the most massive progenitor); if, on the other hand, a halo forms from the merger of two or more equal-mass progenitors then the potential of the new halo will initially be chaotic and the orbits of any galaxies present will be scrambled.

I define two scenarios, which I will call *major* and *minor* mergers. A major merger

occurs when a halo is incorporated into a much more massive halo, while the term minor merger applies to the opposite situation in which a halo grows by a small amount. The critical mass ratio dividing these two regimes is $f_{\rm major}$, the maximum mass ratio for a major merger (note that both terms can describe the *same* merger event; for instance, in a binary merger, the galaxies in the smaller halo might experience a major merger, while those in the larger halo experienced a minor merger).

Major merger If $M_{\rm prev} < f_{\rm major} M_{\rm halo}$, where $M_{\rm halo}$ is the mass of the new halo and $M_{\rm prev}$ the mass of the halo in which the galaxy previously resided, then the galaxy is assumed to enter the new halo on a random orbit. In practice, this means that a new merger timescale is calculated for this galaxy using equation (2.20), based on the properties of the new halo. The merger 'clock' is reset, so that the time available for the galaxy to merge is simply the lifetime of the new halo.

Minor merger If $M_{\rm prev} \geqslant f_{\rm major} M_{\rm halo}$ then I assume that the orbit of the galaxy is essentially unperturbed. In this case, the galaxy retains its merger timescale from the previous halo in which it resided, but the lifetime of the new halo is added to the time available for merging.

2.3.2 Thermal and chemical evolution of gas in haloes

GAS COOLING

Once a halo has virialised, we need to follow the evolution of the baryonic component of the galaxy, initially in the form of hot ($\sim 10^6$ K) gas. As it cools, it loses its pressure support and falls to the bottom of the halo potential well. A galaxy halo begins life with a hot gas reservoir containing a mass $M_{\rm h}$ of hot gas. It is generally assumed that, during virialisation, the hot gas is shock heated to the virial temperature of the halo:

$$T_{\rm gas} = T_{\rm vir} = \frac{\mu m_{\rm p}}{2k} V_{\rm c}^2. \tag{2.21}$$

Hot, ionised gas cools radiatively at a rate, per unit volume, given by

$$L = n_{\rm e}^2 \Lambda(T_{\rm gas}, Z_{\rm gas}), \tag{2.22}$$

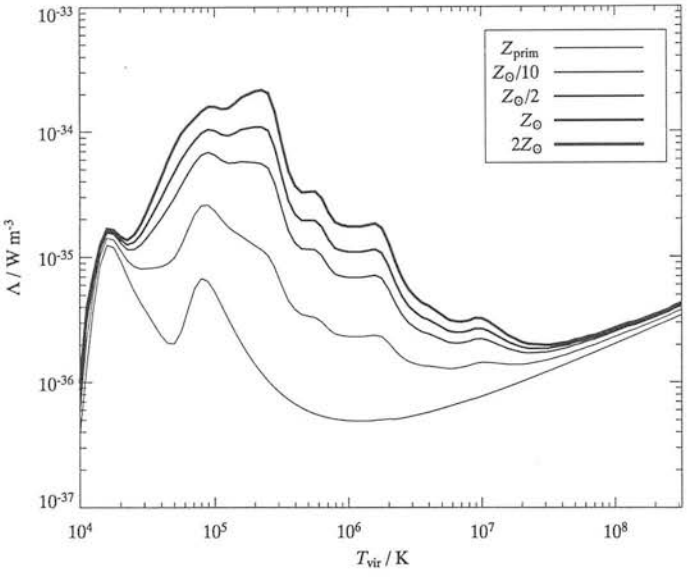

Figure 2.13: The normalised cooling rate of ionised gas as a function of temperature and metallicity (the cooling function). Cooling rates for different metallicities are calculated by linear interpolation of the cooling functions for solar and primordial metallicities, given by Sutherland & Dopita (1993).

where n_e is the electron number density and $\Lambda(T, Z)$ is the cooling function, the value of which depends on the temperature and metallicity of the gas. In figure 2.13 I plot the cooling function for a range of metallicities; these have been calculated by linear interpolation (or extrapolation for $Z > Z_\odot$) of the solar and primordial metallicity cooling functions of Sutherland & Dopita (1993).

We can define a cooling timescale, $\tau_{\rm cool}$, by

$$n_e^2 \Lambda(T_{\rm gas}, Z_{\rm gas}) \tau_{\rm cool} = \frac{3}{2} n k T_{\rm gas}, \tag{2.23}$$

the time taken for the gas to radiate away most of its original energy. Re-arranging equation (2.23) gives

$$\tau_{\rm cool}(\rho_{\rm gas}, T_{\rm gas}) = \frac{3}{2} \frac{\rho_{\rm gas}}{\mu m_{\rm p}} \frac{k T_{\rm gas}}{n_e^2 \Lambda(T_{\rm gas}, Z_{\rm gas})}, \tag{2.24}$$

where $\rho_{\rm gas}$ is the local gas density and $\mu \simeq 0.58$ is the mean molecular weight (assuming a mix of 77% hydrogen and 23% helium). It is assumed that the baryons initially have

an isothermal distribution:

$$\rho_{\rm gas}(r) = \frac{f_{\rm g} V_{\rm c}^2}{4\pi G r^2}, \tag{2.25}$$

where $f_{\rm g}$ is the fraction of the halo mass in the form of hot gas. Hence, the cooling timescale is an increasing function of radius and there will be a radius at which it is equal to the lifetime of the halo (defined to be the time since the formation of the halo or since the last merger event). The mass of gas that can cool over the current lifetime, $t_{\rm life}$, of the halo is thus given as a function of time by the mass of gas contained within this radius. Inserting equations (2.21) and (2.25) into equation (2.24) and setting $\tau_{\rm cool} = t_{\rm life}$ gives

$$r_{\rm cool}(t_{\rm life}) = \left[\frac{\chi_{\rm e}^2 f_{\rm g} \Lambda t_{\rm life}}{3\pi G(\mu m_{\rm p})^2}\right]^{1/2}, \tag{2.26}$$

and hence

$$\begin{aligned} M_{\rm cool}(t_{\rm life}) &= \frac{f_{\rm g} V_{\rm c}^2 r_{\rm cool}}{G} \\ &= \frac{\chi_{\rm e}}{\mu m_{\rm p}(3\pi G^3)^{1/2}} f_{\rm g}^{3/2} V_{\rm c}^2 \Lambda^{1/2} t_{\rm life}^{1/2}, \end{aligned} \tag{2.27}$$

where $\chi_{\rm e} \simeq 0.52$ is the number of electrons per particle, again assuming a 77/23 mix of hydrogen and helium.

Stellar feedback

Gas in the cold interstellar medium can be reheated, and even completely ejected from a halo, by energy output from supernovae and winds associated with young, massive stars. I model this feedback process using the simple parameterisation of CLBF, in which the rate at which gas is reheated is proportional to the star formation rate, $\psi_{\rm SF}$:

$$\dot{M}_{\rm rh} = \beta(V_{\rm c})\psi_{\rm SF}. \tag{2.28}$$

The dependence of the feedback efficiency on the depth of the halo potential well is approximated by a power-law function of the (maximum) circular velocity of the disc:

$$\beta(V_{\rm c}) = (V_{\rm c}/V_{\rm hot})^{-\alpha_{\rm hot}}. \tag{2.29}$$

For the simple case of a massless disc in a halo with an isothermal sphere profile, the circular velocity of the disc is equal to the virial velocity, $V_{\rm c}$, of the halo. The parameter

$V_{\rm hot}$ controls the overall strength of stellar feedback, while $\alpha_{\rm hot}$ controls the dependence on the circular velocity, which is a measure of the depth of the gravitational potential.

The cold gas reservoir is thus depleted at a rate

$$-\dot{M}_{\rm c} = \dot{M}_{\star} + \dot{M}_{\rm rh} - \dot{M}_{\rm cool} \tag{2.30}$$

Reheated gas is added to the hot gas reservoir but is not allowed to cool, or to affect the cooling properties of the hot gas, until after the next merger event. It is neccessary to delay cooling of the reheated gas for feedback to have the desired effect of suppressing star formation. Delaying cooling until after the next merger is an arbitrary choice made mainly for convenience since, in any case, we do not know the distribution, and hence the cooling properties, of reheated gas; after a merger event I assume that the whole of the hot gas reservoir is reheated to the virial temperature of the new halo and adopts an isothermal profile (equation 2.25).

Chemical evolution

The material returned to the cold ISM by stellar winds and supernovae will have been chemically enriched by the nuclear processes inside the stars. Each unit mass of stars that form eject, as part of the recycled gas, a mass y of new metals, where y is the *effective yield*. These new metals are assumed to mix instantaneously with the cold gas. The evolution of the masses of metals, $m_{\rm h}$ and $m_{\rm c}$, in the hot and cold reservoirs respectively is given by

$$\dot{m}_{\rm c} = Z_{\rm h}\dot{M}_{\rm cool} - Z_{\rm c}(1+\beta-R)\psi_{\rm SF} + y\psi_{\rm SF}, \tag{2.31}$$

$$\dot{m}_{\rm h} = -Z_{\rm h}\dot{M}_{\rm cool} + Z_{\rm c}\beta\psi_{\rm SF} + Z_{\rm prim}\dot{M}_{\rm new}, \tag{2.32}$$

where $Z_{\rm c} = m_{\rm c}/M_{\rm c}$ and $Z_{\rm h} = m_{\rm c}/M_{\rm h}$. $\dot{M}_{\rm new}$ is the rate of accretion of new baryons by the galaxy halo; these have primordial metallicity $Z_{\rm prim} \simeq 0.0002$.

The choice to add the metals to the cold gas is one of two extreme possibilities; the alternative scenario is one in which metals are ejected directly into the hot gas, along with gas reheated by supernovae. The real situation is almost certainly some combination of these two extremes: metals are initially ejected into the cold ISM but, since the primary process responsible for this (supernovae) is the same one responsible

for reheating cold gas to the hot phase, it is likely that the reheated gas will have higher than average metallicity. Kauffmann & Charlot (1998) introduce a free parameter to describe the fraction of metals retained in the disc; for simplicity, I assume that the gas reheated by supernovae has a metallicity typical of the disc as a whole.

STRIPPING OF GALAXIES IN MERGED HALOES

Satellites orbiting in a halo have their hot gas reservoirs stripped away and added to the hot gas reservoir of the parent halo, from where it can cool only onto the central galaxy. Thus, satellite galaxies do not accrete any new gas once they are incorporated into a larger halo. They do, however, retain their own reservoirs of cold gas and so can continue forming stars for some time after the merger event. Any gas reheated by star formation is also removed and added to the hot reservoir of the central galaxy.

2.3.3 Star formation in discs

Cooling gas falls to the bottom of the halo potential well, where it forms a thin, rotating disc in centrifugal equilibrium. Stars form in the disc at a rate $\psi_{\rm SF}$, given by

$$\psi_{\rm SF} = \epsilon_\star \frac{M_{\rm c}}{\tau_{\rm disc}}. \tag{2.33}$$

Simple models of star formation, in which the star formation rate (SFR) depends only on local physics, often lead to a star formation law of the form $\psi_{\rm SF} \propto 1/\tau_{\rm disc}$ (e.g. Silk 1997) or $\Sigma_{\rm SF} \propto \Sigma_{\rm gas}^n$ with $n \simeq 1.5$ (e.g. Dopita & Ryder 1994; Sleath & Alexander 1995). This form was also found by Kennicutt (1998) to give a good fit to the observed global SFRs of galaxies under a wide range of conditions, from normal spirals to circumnuclear starbursts. Here, $M_{\rm c}$ is the mass of cold gas available and $\tau_{\rm disc}$ is the dynamical timescale of the disc, which I define to be the orbital time at the disc half-mass radius $r_{\rm half}$:

$$\tau_{\rm disc} \equiv 2\pi R_{\rm half}/V_{\rm c}(R_{\rm half}) \tag{2.34}$$

(this is a different definition to that used by Kennicutt, who defined it at the outer edge of the optical disc; it is, however, the definition preferred by semi-analytical modellers, e.g. CLBF, and I use it to enable direct comparisons with other models). The

dimensionless star formation efficiency, $\epsilon_\star$, is simply the fraction of the available gas converted to stars per dynamical time.

I make the standard assumption of an exponential surface density profile for the cold gas:

$$\Sigma_{\rm c}(R) = \frac{M_{\rm c}}{2\pi R_{\rm d}^2} \exp\left(-R/R_{\rm d}\right), \tag{2.35}$$

where $M_{\rm c}$ is the mass of gas in the cold reservoir. The exponential scale length, $R_{\rm d} = R_{\rm half}/1.678$, is determined by the balance between gravity and the centrifugal force generated by the angular momentum of the disc. For a disc with a flat rotation curve

$$R_{\rm d} = \frac{j_{\rm disc}}{2V_{\rm c}}, \tag{2.36}$$

where $j_{\rm disc}$ is the specific angular momentum of the disc material, which I assume for the moment to be the same as that of the halo, and $V_{\rm c}$ is the (constant) circular velocity. I leave unanswered the question of how the exponential surface mass density profile arises and simply assume that, somehow, the disc material arranges itself into the required configuration and maintains its exponential profile at all subsequent times.

A fraction $\mathcal{R}$ of the mass of stars formed is returned to the cold gas reservoir by stellar winds and supernovae. The mass of long-lived stars thus increases at a rate

$$\dot{M}_\star = (1-\mathcal{R})\psi_{\rm SF}. \tag{2.37}$$

The 'recycled' fraction depends on the adopted IMF (see section 2.3.5) and is a function of the age the stars. For simplicity, I assume that all of the gas is returned instantaneously at the time when the stars form; in this case a value of $\mathcal{R} \simeq 0.3$ is appropriate for a Salpeter IMF. The gas in the cold reservoir is thus used up at a rate

$$\begin{aligned} -\dot{M}_{\rm c} &= \dot{M}_\star + \dot{M}_{\rm rh} - \dot{M}_{\rm cool} \\ &= (1+\beta-\mathcal{R})\psi_{\rm SF} - \dot{M}_{\rm cool}. \end{aligned} \tag{2.38}$$

Integrating equation (2.38) over a time step Δt and assuming that all of the gas cools at the start of the time step gives the mass of stars formed:

$$M_{\rm SF} = \frac{M_{\rm c}^0}{1+\beta-\mathcal{R}} \times \left[1 - {\rm e}^{-(1+\beta-\mathcal{R})\epsilon_\star \frac{\Delta t}{\tau_{\rm dyn}}}\right], \tag{2.39}$$

where $M_{\rm c}^0$ is the mass of cold gas at the start of the time step.

To predict the photometric properties of stellar populations, it is also neccessary to know their metallicities. I assume that the initial metallicity of stars is that of the cold gas from which they formed.

2.3.4 Starbursts

In addition to the quiescent mode of star formation described above, I consider a bursting mode of star formation, triggered by major merger events. A galaxy entering a halo more than $1/f_{\rm major}$ times the mass of the halo in which it previously resided experiences a burst of star formation. Such bursts have two causes: compression of interstellar gas clouds in galaxies as they fall into the higher pressure intracluster medium of the new halo (Evrard 1991) and the increased frequency of interactions, as a result of the new dynamical environment (Moore, Lake & Katz 1998). Such behaviour can be observed in high-redshift clusters in the form of a large population of blue galaxies (Butcher & Oemler 1978).

I approximate the duration of the burst by an estimate of the dynamical timescale of the halo:

$$t_{\rm burst} = \frac{r_{\rm half}}{V_{\rm c}}, \tag{2.40}$$

where $r_{\rm half}$ is the half-mass radius of the halo. During this time, hot gas continues to cool onto the most massive galaxy, increasing the amount of cold gas available to fuel the starburst. Quiescent star formation in the discs of galaxies is suppressed until after the bursting period is over. I model the starburst itself using the simple scaling law of Cole et al. (1994):

$$\tau_{\rm burst}(V_{\rm c}) = \tau^0_{\rm burst}\left(\frac{V_{\rm c}}{300\ {\rm kms}^{-1}}\right)^{\alpha_{\rm burst}}. \tag{2.41}$$

It is often assumed that the increased frequency of galaxy–galaxy interactions results in efficient channeling of gas into the centres of galaxies, triggering starburst activity in the nuclei, but the faint blue galaxies seen in high-redshift clusters are of predominantly late types (Dressler et al. 1994). Moore, Lake & Katz (1998) showed that frequent encounters between galaxies in the cluster environment initially result in distorted spirals, which evolve into ellipticals over timescales of Gyr. For these reasons, I choose to place stars formed in bursts in the discs of galaxies.

2.3.5 Magnitudes and colours

STELLAR SPECTRAL SYNTHESIS

The star formation histories of galaxies are used to predict the magnitudes and colours of galaxies by combining them with the GISSEL library of synthetic stellar populations (Bruzual & Charlot 1993). I interpolate between the spectra provided to generate a grid of simple stellar populations (SSPs) — populations of mass 1 $M_\odot$ formed in an instantaneous burst at some time — with ages at 0.01 Gyr spacings, for a range of metallicities. These are weighted according to the star formation history of a given galaxy and added together to give the present-day spectrum:

$$L_\lambda(t_0) = \sum_i \int_0^{t_0} \psi^i_{\rm SF} L^{\rm SSP}_\lambda(Z^i_*, t_0 - t)\, {\rm d}t. \tag{2.42}$$

It is neccessary to sum over all progenitors, i, of the current galaxy, which potentially consists of many separate stellar populations that formed at the same time, but in different progenitor haloes and hence with different metallicities.

The present-day luminosity, at a given wavelength, of a SSP formed at time t is

$$L^{\rm SSP}_\lambda(Z, t_0 - t) = \int_{m_1}^{m_2} \left.\frac{{\rm d}n}{{\rm d}m}\right|_m l_\lambda(m, Z, t_0 - t)\, {\rm d}m, \tag{2.43}$$

where $l_\lambda(m, Z, t_0 - t)$ is the luminosity of a single star of mass m and metallicity Z formed at time t. The spectrum of a stellar population depends not only on its age, but also on its metallicity; in Fig. 2.14 I plot the evolution with age of 4 broadband magnitudes (B, V, I and K) and 2 colours ($B - V$ and $B - K$) for SSPs with three different metallicities: one solar ($Z_\odot = 0.02$), one sub-solar ($Z = Z_\odot/10$) and another super-solar ($Z = 2.5 Z_\odot$). The most obvious features in Fig. 2.14 are that stellar populations of all metallicities fade over time and become redder. There is a colour degeneracy between age and metallicity, in the sense that, in addition to old stars being redder than young stars, high metallicity stars of a given age are redder than low metallicity ones.

The evolution of a SSP also depends on the IMF. The curves in Fig. 2.14 are for a Salpeter (1955) IMF:

$${\rm d}n/{\rm d}m \propto m^{-(1+x)} \qquad (x = 1.35), \tag{2.44}$$

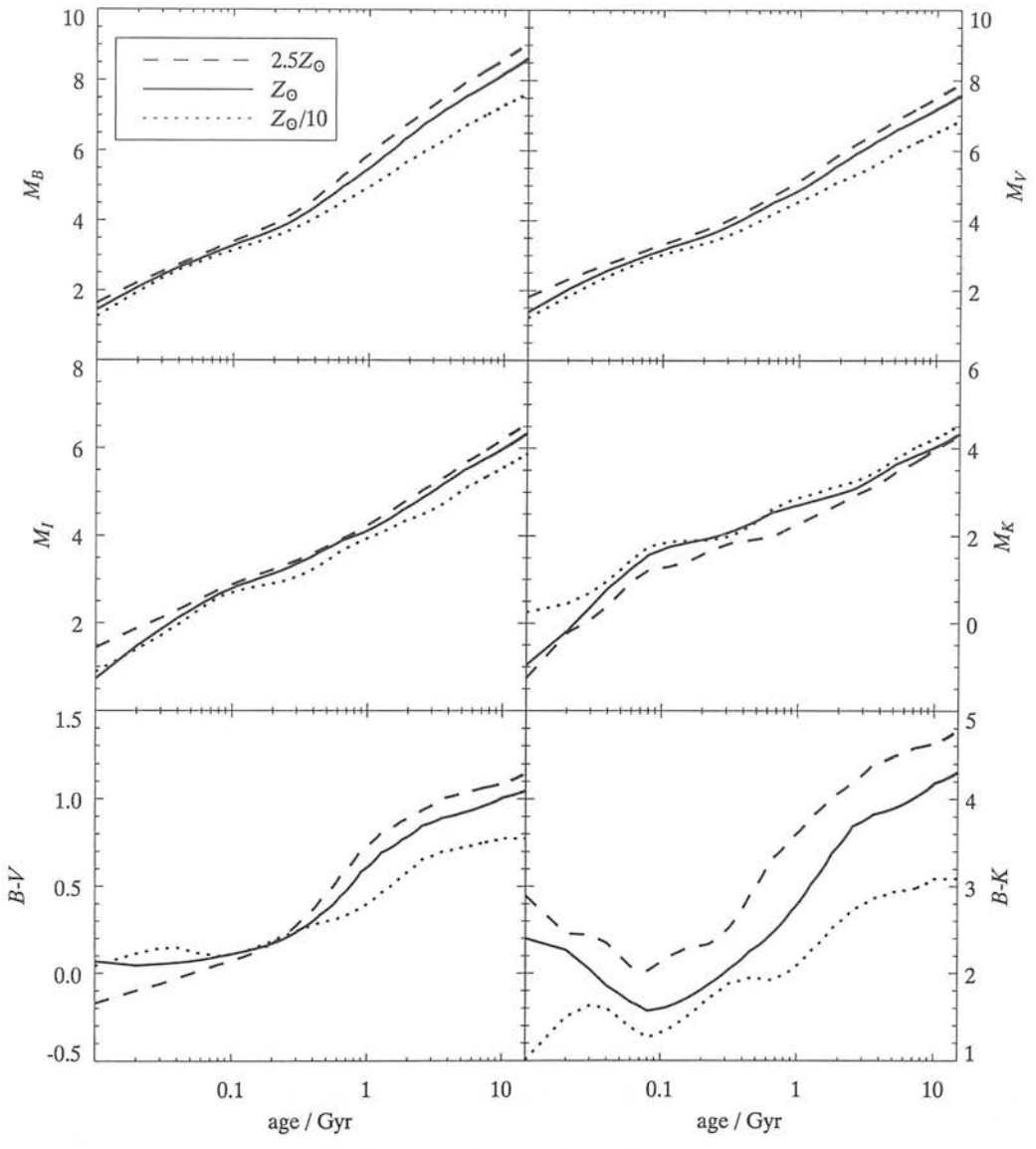

Figure 2.14: Evolution of the broadband colours and magnitudes of a simple stellar population of mass 1 $M_\odot$. The different curves in each panel illustrate the behaviour of stellar populations of different metallicities: solar ($Z_\odot = 0.02$, solid lines), sub-solar ($Z = Z_\odot/10$, dotted lines) and super-solar ($Z = 2.5Z_\odot$, dashed lines).

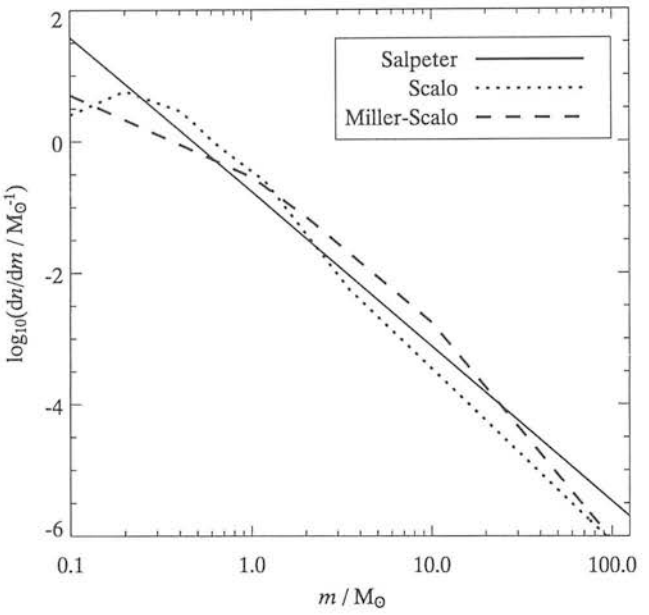

Figure 2.15: The stellar initial mass functions of Salpeter (1955, solid line), Scalo (1986, dotted line) and Miller & Scalo (1979, dashed line). Each curve is normalised to a total mass of 1 $M_\odot$ between upper and lower limits of 125 $M_\odot$ and 0.1 $M_\odot$, respectively.

which I assume throughout this thesis. The GISSEL library also provides synthetic spectra for two other choices of IMF: those of Scalo (1986) and Miller & Scalo (1979); these take the form of broken power laws with several segments (6 in the case of Scalo and 4 for Miller-Scalo). In Fig. 2.15 I compare the three posible choices of IMF. All three curves are normalised to a total mass of 1 $M_\odot$. Qualitatively, all three IMFs have the same behaviour: the number of stars exhibits a power-law increase with decreasing mass, down to the limit of 0.1 $M_\odot$. Stars below this mass do not have a high enough central pressure to trigger widespread nuclear fusion; the numbers of such faint 'brown dwarfs' is uncertain and can be allowed for by introducing the parameter $\Upsilon = M_\star(\mathrm{total})/M_\star(> 0.1\ M_\odot)$. The main consequence of a large population of such stars, if it did exist, would be to lock up some fraction of the available baryons in a dark form. Note that they should not be included in the star formation law, since this is motivated by observations which also rely on an assumed IMF and implicitly neglect the brown dwarf population (e.g. Kennicutt, Tamblyn & Congdon 1994). The

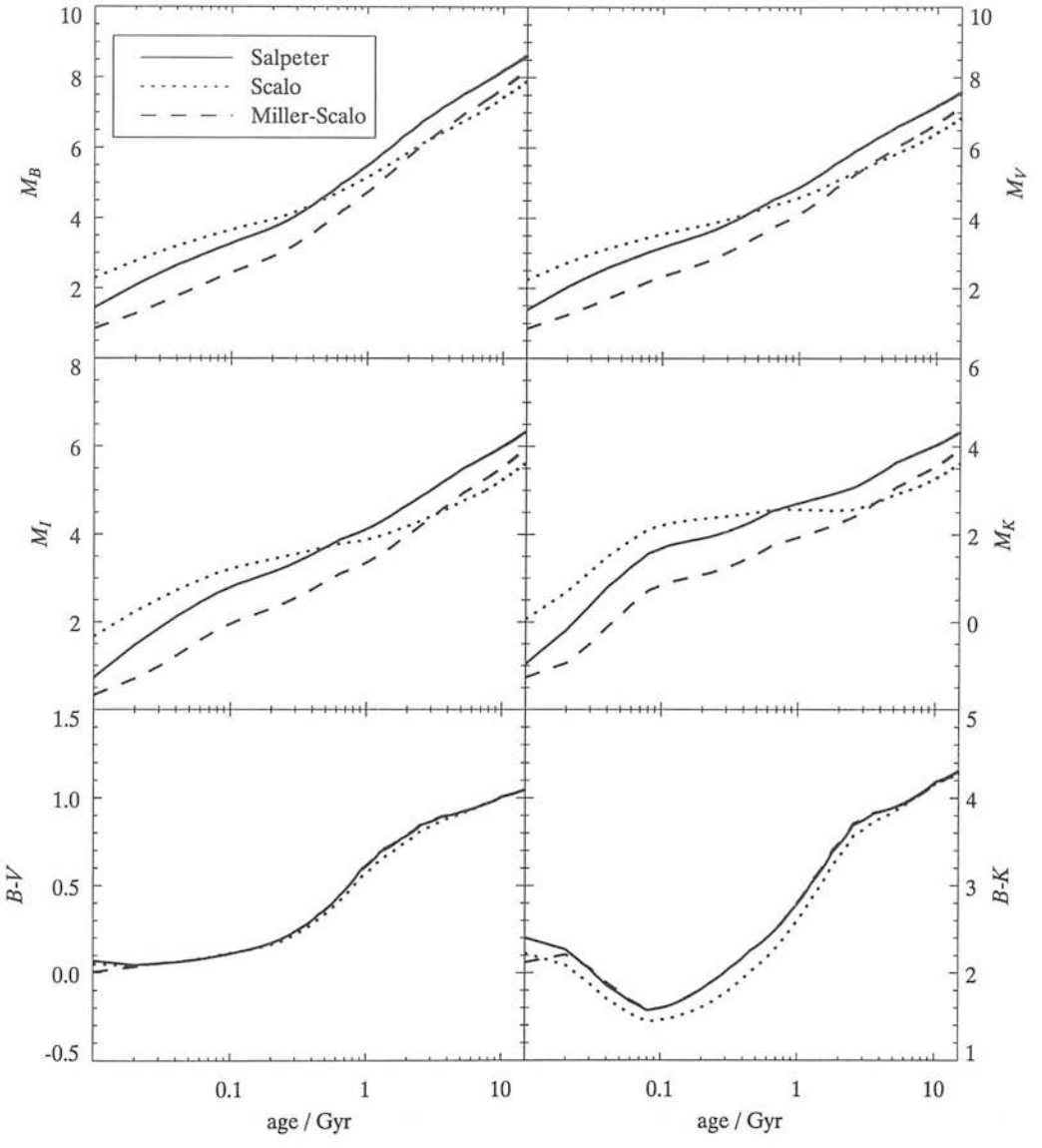

Figure 2.16: Evolution of the broadband colours and magnitudes of a simple stellar population of mass 1 $M_\odot$. The different curves in each panel are for different initial mass functions: Salpeter (1955, as used in this thesis, solid line), Scalo (1986, dotted line) and Miller & Scalo (1979, dashed line).

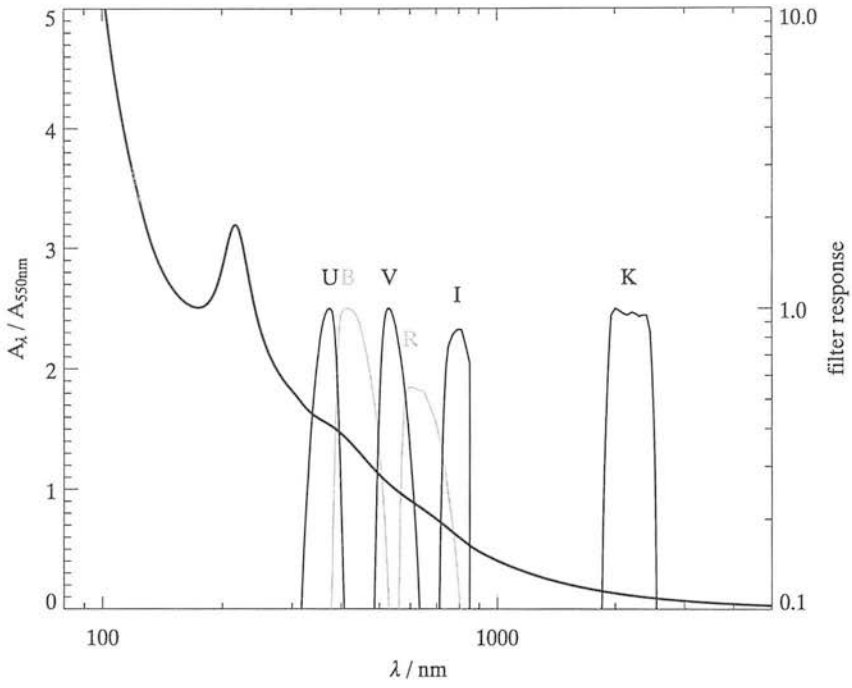

Figure 2.17: The Galactic interstellar extinction law of Cardelli, Clayton & Mathis (1989). A_λ is the average amount of extinction, in magnitudes, at wavelength λ and the curve has been normalised to unity at a wavelength of 550 nm, corresponding to the centre of the V band. Superimposed are the responses of some of the standard Johnson filters, in logarithmic units.

evolution of the broadband magnitudes and colours of SSPs with the same mass (1 $M_\odot$) and metallicity (solar), but different IMFs, is illustrated in Fig. 2.16. Although the integrated flux in a given waveband depends strongly on the assumed IMF, the colours are essentially independent. This is because, at a given age, the stellar populations are dominated by the same type of star — those close to the main sequence turn-off — and hence have the same spectral shape; the only difference is in the normalisation, which is determined by the number of those stars. For young populations, the Salpeter IMF lies between the other two but, for old populations, it gives more flux because of the larger numbers of low-mass stars.

DUST ABSORPTION

Stellar spectral synthesis models give us the spectra of 'naked' stellar populations. To reproduce the observed spectra of galaxies, we need to take into account the light which is absorbed by dust within the galaxy. The physics of dust formation and its distribution in galaxies is poorly understood and, consequently, difficult to model directly. For this reason, I adopt a simple model, in which the amount of extinction scales with the luminosity of the disc. Wang & Heckman (1996) find the following relation between the face-on optical depths of disc galaxies and their absolute luminosities in the B band:

$$\tau_B = \tau_B^\star \left(\frac{L_B}{L_B^\star} \right)^\beta . \tag{2.45}$$

Here, $L_B^\star$ is the unextincted B-band luminosity corresponding to the knee of the Schechter function and $\tau_B^\star$ is the typical optical depth of an $L_B^\star$ galaxy. I adopt values of $\tau_B^\star = 0.8$, $L_B^\star = 1.3 \times 10^{10}$ (corresponding to an intrinsic magnitude of $M_B^\star = -20.5$) and $\beta = 0.5$, as found by Wang & Heckman.

For a 1D slab, with face-on optical depth τ_B in the B band, the extinction is a function of inclination, i, and is given by

$$A_B = -2.5 \log \left(\frac{1 - \mathrm{e}^{-\tau_B \sec i}}{\tau_B \sec i} \right) . \tag{2.46}$$

I determine the value of τ_B for an individual galaxy using observations of local spiral galaxies. I assume that dust absorption affects only light from the disc. As yet, I take no account of the chemical evolution of galaxies when calculating the optical depth to dust absorption. Optical depths in the B band are converted to other bands using the mean Galactic extinction spectrum of Cardelli, Clayton & Mathis (1989), which is shown in Fig. 2.17. The optical depth at wavelength λ is given by

$$\tau_\lambda = \tau_B (A_\lambda / A_B)_{\mathrm{Gal}}, \tag{2.47}$$

where A_λ is the mean extinction at wavelength λ measured in the Galaxy.

2.3.6 Dealing with finite mass resolution

A minimum of 10 particles is set on the size of groups found by the percolation algorithm. This corresponds to a halo mass of $\sim 10^{10}$ $\mathrm{M}_\odot$. In the standard picture of

hierarchical structure formation, it is expected that gas will be able to cool at high redshifts in structures smaller than this minimum mass. These structures will eventually be incorporated into the larger haloes seen in the N-body simulation, along with any stars and cold gas they contain. As well as allowing more time for gas to cool, 'old' stellar populations have two effects on the evolution of galaxies at later times: they lock up cold gas, making it unavailable for *in situ* star formation in larger haloes and they pre-enrich the remaining gas with metals, with the result that gas cools faster when it is finally incorporated into a resolved halo and later stellar populations are redder. These effects are expected to be small in most cases (an assumption that will be tested later), particularly since feedback is very efficient at suppressing star formation in low-mass haloes, but I attempt to account for them using a simple model.

When a halo accretes new mass, which has not previously been in a resolved halo, I assume that it is in the form of identical sub-resolution haloes with masses $M_{\rm min}/2$, where $M_{\rm min}$ is the minimum halo mass imposed by the halo finder. I adopt a formation time of $t_{\rm form}/2$ for the old populations, where $t_{\rm form}$ is the formation time of the first resolved structure into which the matter is incorporated, so that the time available for gas to cool is $t_{\rm old} = t_{\rm form}/2$. The virial velocity of a halo of mass M at time t scales approximately as $V_{\rm c}^3 \propto M/t$ (this is exact only for the case of a flat matter-dominated universe; however, at redshifts $z > 1$, the contribution from vacuum energy is negligible, so this is a good enough approximation, given the simplicity of the model). This yields

$$V_{\rm old} \simeq V_{\rm c} \left(\frac{M_{\rm min}}{M_{\rm halo}} \right)^{1/3}, \tag{2.48}$$

where $V_{\rm c}$ is the virial velocity of the resolved halo. Since this part of the modelling does not need to be accurate, these approximations should be sufficiently good.

The rate of gas cooling in sub-resolution haloes is given by the equivalent of equation (2.24):

$$M_{\rm cool}(t_{\rm old}) = \frac{\lambda e}{\mu m_{\rm p} (3\pi G^3)^{1/2}} f_{\rm g}^{3/2} V_{\rm old}^2 \Lambda^{1/2} t_{\rm old}^{1/2}. \tag{2.49}$$

I assume primordial metallicity to calculate $\Lambda(T, Z)$.

Rather than calculate the star formation timescale from the disc properties, I model it as a simple power law of the circular velocity of the haloes:

$$\tau_* = \tau_{*\,({\rm old})}^0 \left(\frac{V_{\rm old}}{300 \ {\rm km \ s^{-1}}} \right)^{\alpha_{*\,({\rm old})}}. \tag{2.50}$$

This simple star formation prescription was used by Cole et al. (1994) to model star formation in galaxies of all sizes and provided a good fit to many of the observed properties of galaxies. By using this parameterisation I also avoid making any assumptions about the properties of the discs (if indeed any exist) that reside within these small haloes. The old populations are added to the bulge of the central galaxy as if they were merging satellites and any remaining cold gas is added to the disc of the galaxy.

The values chosen for the parameters $V_{\rm old}$ and $t_{\rm old}$ are only approximations to the average properties of the sub-resolution haloes. Although stellar populations formed in these structures make up only a small fraction of the present-day light of galaxies, they become increasingly important at higher redshifts. It will be neccessary, as the other ingredients of the models improve, to model this aspect of galaxy formation in a more realistic way. A more satisfactory approach would be to extend the merger tree to lower mass haloes, either using higher resolution simulations or by attaching analytic 'leaves' to the ends of the merger tree.

2.4 Constraining the model: a basic parameter set

The parameters of the galaxy formation model are summarised in table 2.3, where I also list the values adopted in the basic parameter set. The time step, Δt, over which gas cooling, star formation, feedback and chemical evolution are evaluated is chosen to be much smaller than the time resolution of the merger trees. In principle, steps as small as 10 Myr are possible; this is the age of the youngest SSP. As in vKJP, I take the properties of the stellar populations to be fixed in order to investigate the effects of varying the galaxy formation recipe. I adopt a Salpeter (1955) IMF, and fix the contribution from low-mass stars (Υ) to give the observed number of $L^\star$ galaxies for the basic parameter set. It is tempting to assume $\Upsilon = 1$, as in vKJP, but this seems to be inconsistent with the observations, as I will show. The effective stellar yield and recycled fraction are within the bounds set by stellar evolution models (e.g. Woosley & Weaver 1995; Marigo 2001). I also fix the parameters of the dust model, adopting the values given by Wang & Heckman (1996). In the following sections I investigate the consequences of varying each of the remaining parameters in turn on the luminosities

Table 2.3: Free and fixed parameters of the model and the values adopted in the basic set.

Description	Parameter	Value
GENERAL PARAMETERS		
Time step	Δt	0.04 Gyr
FREE PARAMETERS		
Galaxy merging within haloes:		
dynamical friction time-scale	$f_{\rm mrg}$	0.5
Star formation in discs:		
star formation efficiency	$\epsilon_\star$	0.05
Bursting star formation:		
bursting mode star formation timescale	$\tau^0_{\rm burst}$	0.1 Gyr
bursting mode star formation scaling law	$\alpha_{\rm burst}$	-1.5
Feedback:		
feedback efficiency	$V_{\rm hot}$	200 km s^{-1}
feedback scaling law	$\alpha_{\rm hot}$	2.0
Sub-resolution star formation:		
star formation timescale	$\tau_{\star\,({\rm old})}$	2.0 Gyr
star formation scaling law	$\alpha_{\star\,({\rm old})}$	-1.5
FIXED PARAMETERS		
Stellar populations:		
initial mass function	IMF	Salpeter
ratio of all stars to luminous stars	Υ	1.3†
effective stellar yield	y	0.02
recycled fraction	$\mathcal{R}$	0.3
Dust extinction:		
dust extinction normalisation	$\tau^\star_B$	0.8
characteristic luminosity	$L^\star_B$	1.3×10^{10} L$_\odot$
dust extinction scaling law	β	0.5

† Fixed to give a good match to the B-band LF at $M^\star$

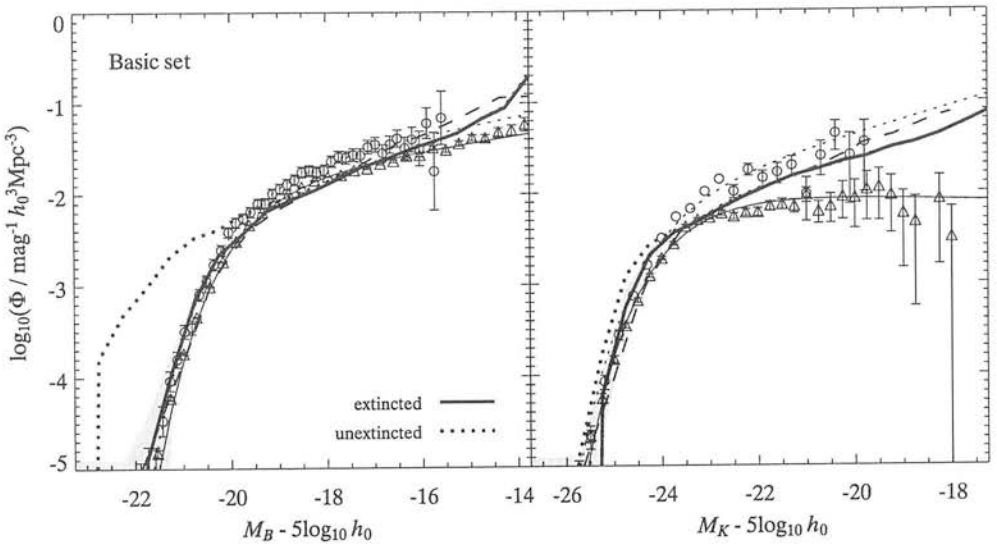

Figure 2.18: B- (top row) and K- (bottom row) band luminosity functions (LFs) for the basic parameter set. The heavy, solid line in each panel is the LF of the model galaxies and the grey shaded area marks the r.m.s. variance between the different boxes. This is plotted over recent observational data as follows: in the B band from Norberg et al. (2002, triangles & solid line) and Blanton et al. (2001, circles & dotted line); in the K band from Cole et al. (2001, triangles & solid line), Huang et al. (2003, circles & dotted line, error bars are shown where these were readable in the published plot). I also plot the LFs for the reference model of CLBF (dashed line).

of galaxies.

2.4.1 Luminosity functions

The model B- and K-band LFs predicted for the basic parameter set are shown in Fig. 2.18. The heavy, solid curve is the mean LF over the five simulations and the shaded area marks the r.m.s. variance between the simulations. Note that this is an underestimate of the true magnitude of cosmic variance, since it does not take into account clustering — all of the boxes used here are constrained to have mean density equal to the universal value whereas, in reality, bright galaxies are strongly clustered and follow the large-scale structure. Cosmic variance is also a major source of uncertainty in observational measurements of the LF and may account for some of the differences between the results of different surveys (Cross et al. 2001). I compare to the two most recent measurements of the LF in each band: those of Norberg et al. (2002) and Blanton

et al. (2001) in B and Cole et al. (2001) and Huang et al. (2003) in K. I also plot the reference model of CLBF; this is the dashed curve in Fig. 2.18. I also plot the LF for the same parameters using unextincted magnitudes (heavy, dotted curves). Given the difference this makes to the bright end of the LF, it is surprising that early models (Cole et al. 1994; vKJP), which neglected dust extinction, were able to come close to matching the bright end of the B-band LF. Even in the K band, which is relatively unaffected by dust, there is a considerable degree of extinction in the brightest galaxies. My estimates of the amount of dust extinction in galaxies is observationally motivated and should be fairly robust but, along with uncertainties about the IMF and stellar evolution, the exact nature of dust extinction contributes to the overall uncertainty in the luminosities of the model galaxies.

In spite of these uncertainties, the bright end of the model LF fits the data extremely well, particularly at optical wavelengths. This is a major achievement, given the simplicity of the galaxy formation model, and was achieved with minimal adjusting of the free parameters. In fact, as I show below, this basic result is reproduced for all reasonable choices of the parameters listed in table 2.3. There is a lack of bright galaxies in the K band on average, indicating that the brightest galaxies are slightly too blue. The most noticeable discrepancy between model and data is at the faint end: the model overpredicts the number of galaxies faintwards of $M_B = -15.5 + 5\log_{10} h_0$ and $M_K = -20 + 5\log_{10} h_0$, if the data in these regions are to be believed. This discrepancy is common to most galaxy formation models, including SAMs. It is a manifestation of the 'dwarf galaxy problem' — the observation that the expected spectrum of density fluctuations in a CDM universe leads to a halo mass function that rises steeply with decreasing mass, while the LF is essentially flat at the faint end. However, if the data of Blanton et al. (2001) are extrapolated to fainter magnitudes then the disgreement in the B band lessens and in the K band there is a significant disagreement between the two most recent estimates of the LF, with both models (mine and that of CLBF) favouring the higher Huang et al. (2003) data. Any attempt to match the faint end of the observed luminosity function is complicated by considerations of completeness in the observational samples. This issue is dealt with in chapter 4.

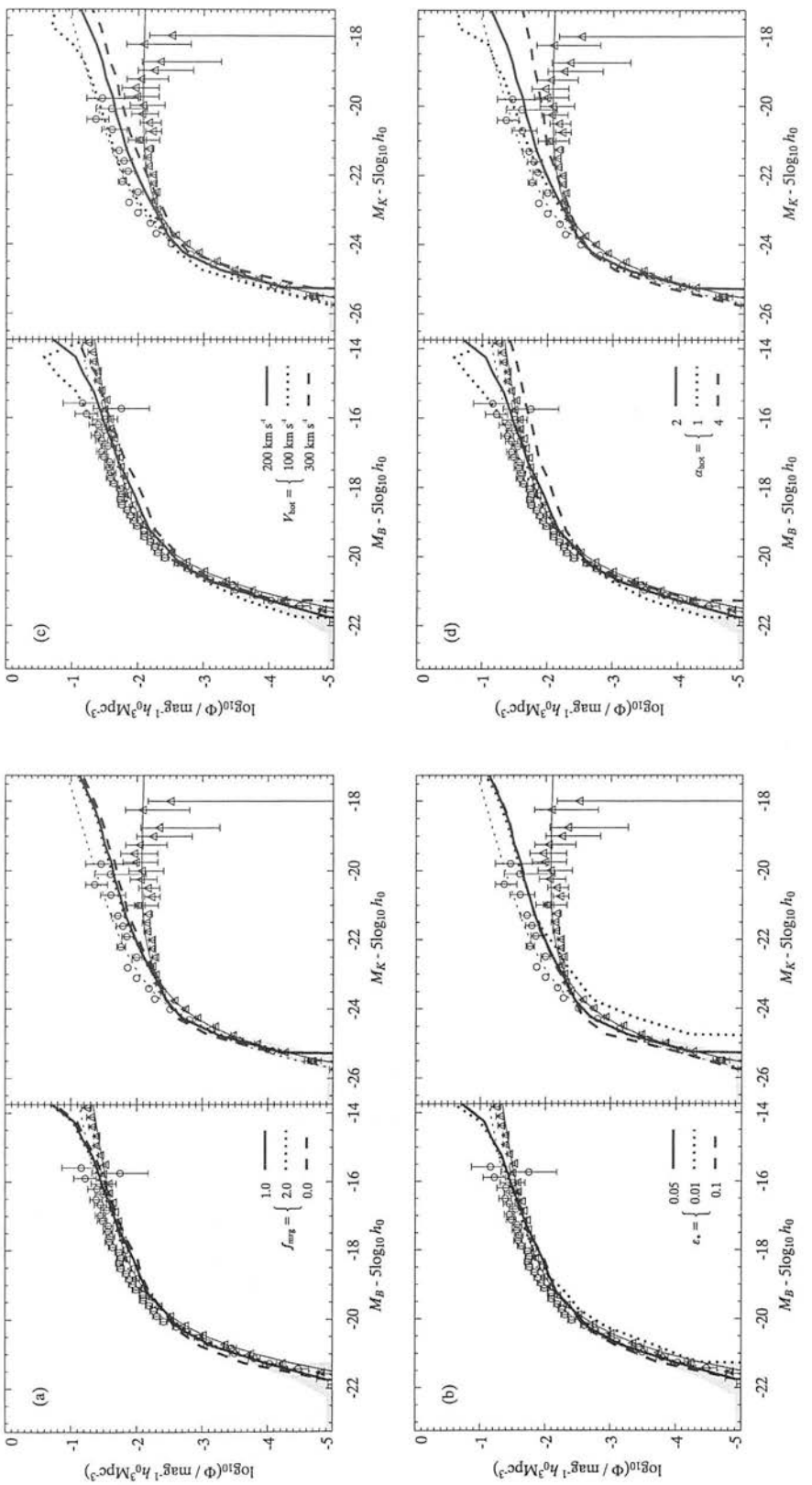

Figure 2.19: The effect on the B- and K-band luminosity functions (LFs) of varying the free parameters of the model: (a) the dynamical friction timescale, $f_{\rm mrg}$; (b) the star formation efficiency in discs, ϵ_*; (c) the feedback efficiency, $V_{\rm hot}$; and (d) the slope of the feedback scaling law, $\alpha_{\rm hot}$. The stellar mass-to-light ratio is fixed at $\Upsilon = 1.3$. The solid curve in each panel is the model LF for the basic parameter set and the shaded area marks the r.m.s. variance between the five simulation boxes. The model LFs are plotted over observational data as in Fig. 2.18.

Varying the free parameters

In Figs. 2.19 and 2.20 I show the effects on the LFs of varying each of the free parameters of the model independently of the others.

Dynamical friction Varying the dynamical friction timescale over a reasonable range has almost no effect on the model LF, as shown in Fig. 2.19(a). This insensitivity to the rate of galaxy merging is a result of the assumption that there will always be a dominant galaxy in every subhalo, onto which hot gas, some of it stripped from the orbiting satellites, can cool. The brightest galaxies are usually the dominant galaxies in their subhaloes and so their luminosities are determined primarily by the rate of gas cooling in these subhaloes. Merging satellites bring in a few stars but little new gas. Instant (or, at least, very rapid) merging in subhaloes, whilst unlikely, actually gives the best fit to the bright end of the B-band LF, while the effect on the bright end of the K-band LF is small and more or less within the errors of the observations. This is a rather nice result, since it means that the approach outlined in section 2.2 is a viable alternative to the empirical dynamical friction argument and suggests that the merging of substructure within haloes really does lie behind the shape of the galaxy luminosity function.

Star formation and feedback Increasing the star formation efficiency leads to more stars being formed. Again, the change in Fig. 2.19(b) is small because, in the long-term, the star formation histories of the brightest galaxies closely follow the gas accretion histories. The largest change is in the K band, which is generally assumed to be a good indication of total stellar mass. On short timescales, the effect of increasing $\epsilon_\star$ is to form more stars earlier, leading to redder populations. This reddening is added to by the fact that metals are synthesised more rapidly so that stars formed later on are also redder, for their age, than they would otherwise have been.

Feedback from star formation is a necessary ingredient in galaxy formation models to produce a flat faint-end slope to the LF. It works by suppressing star formation very effectively in low-mass haloes. The mechanism is generally assumed to be ejection of gas from the central regions by stellar winds and supernovae, which is more efficient

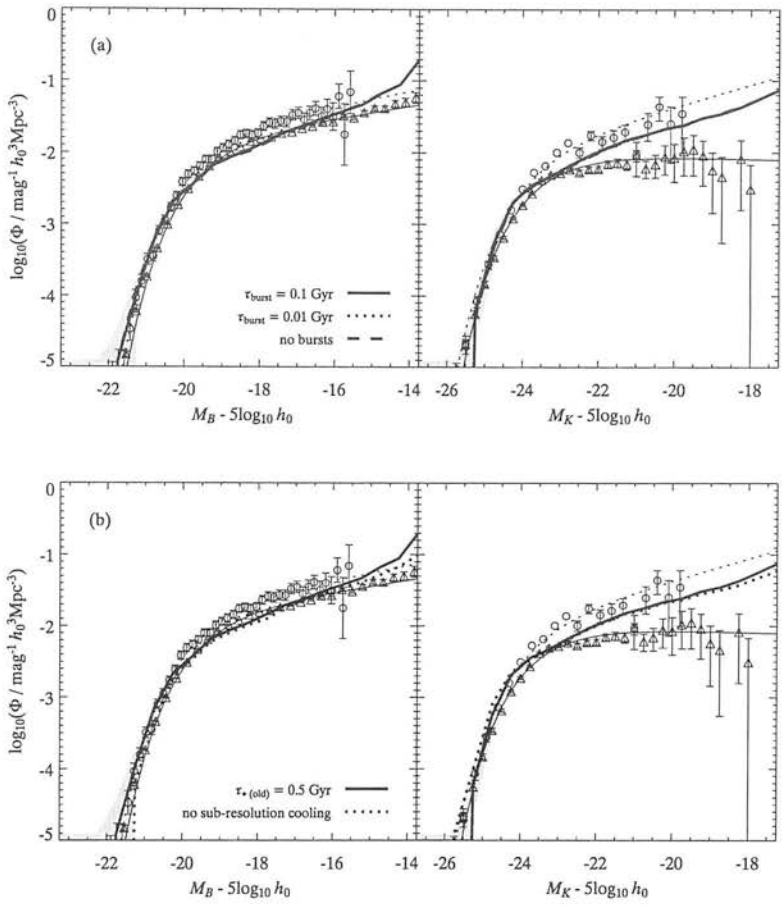

Figure 2.20: As Fig. 2.19 for (a) different starburst strengths and (b) no sub-resolution cooling.

in shallower potential wells. The feedback prescription has two free parameters: $V_{\rm hot}$ determines the overall strength of feedback, while $\alpha_{\rm hot}$ controls the dependence of feedback efficiency on the circular velocity. Increasing either of these parameters (Fig. 2.19c and d) has the desired effect of flattening the faint end of the LF, while having little effect on the brightest galaxies. As found by most previous work in this area, a high value of $\alpha_{\rm hot}$ is neccessary to explain the data faintwards of $M_B \simeq -16 + 5\log_{10} h_0$. However, as noted by CLBF, the faint end of the observed LF is severly affected by selection biases and no definitive measurement of the intrinsic slope exists, so a poor match to the faint end should not be considered a major failing of the model. The issue of selection effects is discussed in more detail in later chapters.

Starbursts One of the novel features of my model is the inclusion of a bursting mode of star formation driven by *halo* merger events (as opposed to bursts on galaxy–galaxy mergers). A strong bursting mode of star formation can give results similar to a traditional monolithic collapse model, but the prescription used here gives only weak bursts that make little difference to the present-day LF (Fig. 2.20a). In Fig. 2.21 I show the relative contributions made by the quiescent (disc) and bursting modes of star formation to the cosmic star formation history. Star formation in discs dominates at all redshifts in this model, even at early times when the halo merger rate is high. Overall, the star formation history is rather insensitive to the mode of star formation and is dependent primarily on the mass accretion rate of the haloes in the simulation.

Sub-resolution star formation Fig. 2.21 shows that, as expected, gas cooling and star formation in haloes below the resolution of the N-body simulation makes a relatively small contribution ($\sim 7\%$) to the present-day stellar content of the universe. This is also clearly reflected in the LFs in Fig. 2.20(b). Gas in sub-resolution haloes is prevented from cooling efficiently by strong stellar feedback and the few stars that formed in these haloes are old and faint. In fact, the model reproduces the observed LF better without this ingredient in place, suggesting that even stronger feedback is needed to prevent significant amounts of star formation in these haloes.

2.4.2 Tully–Fisher relation

Although the Tully–Fisher relation (TFR) is, strictly speaking, only followed by spiral galaxies, it is used as a general constraint on galaxy formation models since spirals make up a large fraction of the bright galaxy population in the local universe. In Fig. 2.22 I plot the model I-band TFR for the basic parameter set. I have made no attempt at morphological selection by, for example, bulge-to-disc ratio, but I have selected only galaxies that are actively forming stars. A galaxy is considered to have active star formation if its star formation rate (not including starbursts), averaged over the last 500 Myr (roughly the time between the last two snapshots of the simulation), is > 0.01 $M_\odot$ yr^{-1}. This cutoff, although rather arbitrary, is typical of the SFRs found in the nuclei of the more active E/S0 galaxies (Sadler et al. 2000). The model

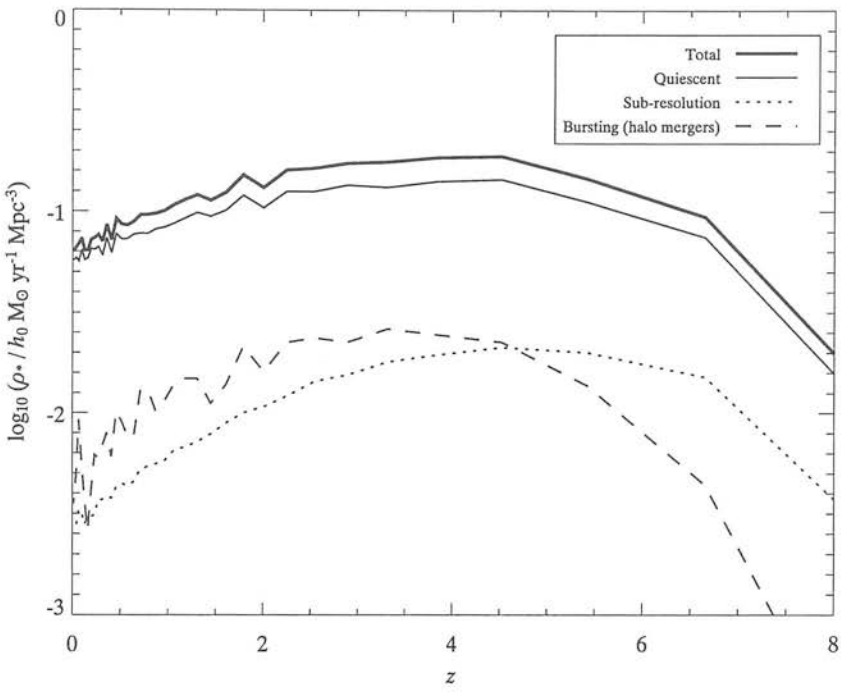

Figure 2.21: The star formation history of the universe as a function of redshift for the basic parameter set. The star formation rate is broken down into the contributions from different modes: quiescent star formation in discs (light, solid line), starbursts (dashed line) and star formation in sub-resolution haloes (dotted line).

TFR is not changed significantly when the cut-off is varied by a factor of ten in either direction. The magnitudes plotted include the effects of dust extinction, assuming the disc to be face-on to the observer (Tully–Fisher samples are usually converted to face-on values using an average correction,which may introduce extra scatter into the observed relation).

The model reproduces the slope *and* zero-point of the observed TFR well, except at the bright end, where it falls below the observed relation. To achieve this, it was necessary to fix the stellar mass-to-light ratio at $\Upsilon = 1.3$. A parameter set with $\Upsilon = 1$ and a lower star formation efficiency, $\epsilon_\star = 0.01$, produces galaxies of the same overall luminosity, but the median luminosity for star-forming galaxies is too low and there are fewer of them. Note that the velocities plotted in Fig. 2.22 are the *virial* velocities of the haloes. Cole et al. (2000) found that when the *disc* circular velocities are used

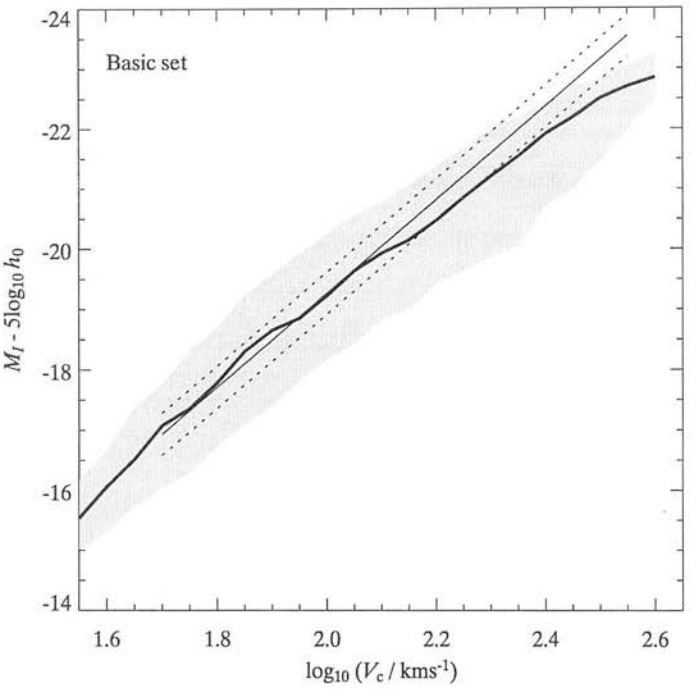

Figure 2.22: The I-band Tully–Fisher relation (TFR) for the basic parameter set. Galaxies are selected from the simulations as described in the text. The heavy solid line is the median relation and the shaded area marks the 10% and 90% quartiles of the distribution. The light, solid line is the mean observed TFR, using four data sets (Pierce & Tully 1992, Mould et al. 1993, Giovanelli et al. 1997, Mathewson, Ford & Buchhorn 1992) and the dotted lines show the typical scatter in the observed relation (Giovanelli et al. 1997).

(taking into account the self-gravity of the baryons and the halo density profile, which may not be isothermal) the observed TFR is no longer well-matched by semi-analytic models. I investigate these effects in chapter 3. In Fig. 2.23 I illustrate the effects of varying some of the other parameters of the model (fixing Υ at 1.3). The main ingredients that affect the TFR are feedback, which determines the luminosities of the lowest circular velocity galaxies, and the efficiency of star formation, which affects galaxies of all masses. Feedback tends to introduce curvature into the TFR; very strong feedback can be ruled out on these grounds, but weaker feedback than assumed in the basic parameter set is allowed.

The TFR for individual observational samples is very tight, having a typical scatter

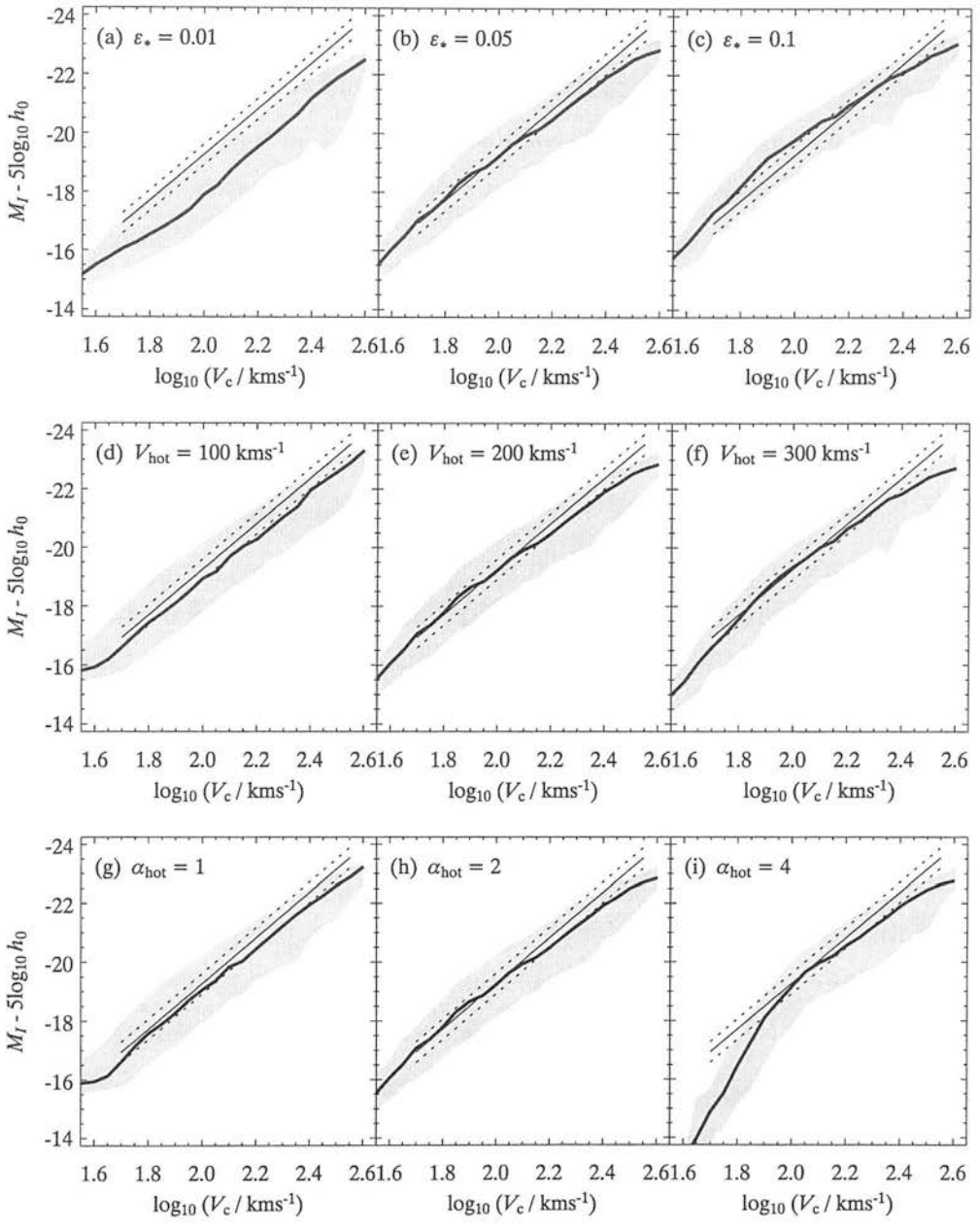

Figure 2.23: The effect on the I-band Tully–Fisher relation (TFR) of varying the free parameters of the model: (a)–(c) the star formation efficiency in discs, $\epsilon_\star$; (d)–(f) the feedback efficiency, $V_{\rm hot}$; and (g)–(i) the slope of the feedback scaling law, $\alpha_{\rm hot}$. The stellar mass-to-light ratio is fixed at $\Upsilon = 1.3$. Galaxies are selected from the simulations as described in the text. The heavy solid line in each panel is the median relation and the shaded areas mark the 10% and 90% quartiles of the distributions. The light, solid line is the mean observed TFR, using four data sets (Pierce & Tully 1992, Mould et al. 1993, Giovanelli et al. 1997, Mathewson, Ford & Buchhorn 1992) and the dotted lines show the typical scatter in the observed relation (Giovanelli et al. 1997).

of around 0.35 I-mag (Giovanelli et al. 1997), although different samples often yield quite different values for the zero-point and, most especially, the slope. A formal, linear fit to the model relation yields a scatter of 0.73 magnitudes for the basic parameter set, roughly twice the observed scatter. This is no doubt partly a result of the scatter in the halo properties measured in the simulations, something that is too often not taken into account in galaxy formation models.

2.5 Discussion: consequences of subhalo merger trees

Much emphasis has, in recent years, been placed on explaining the discrepancy between the faint-end slope of the galaxy LF and the predictions of galaxy formation models. As stated above, this discrepancy stems from the fact that the spectrum of fluctuations in CDM cosmologies has most power on small scales, so the absence of large numbers of faint galaxies might indicate a problem with the underlying cosmological model. However, it is becoming clear that explaining the bright end of the LF is at least as much of a challenge for current galaxy formation models. In the currently favoured picture of galaxy formation, gas falls into potential wells created by gravitational collapse of dark matter fluctuations (e.g. White & Rees 1978; White & Frenk 1991). During the collapse and virialisation of these 'haloes', the gas is shock-heated to the virial temperature of the haloes. Subsequently, if the density is high enough, it cools radiatively and falls to the bottom of the potential wells, where it forms stars. The most massive self-gravitating structures in the universe today are rich clusters of galaxies. These contain large reservoirs of hot gas and, were they to be fully virialised, as assumed in the Press-Schechter formalism, then they would be expected to have cooling times much less than a Hubble time (Fabian 1994). Such objects would be expected to exhibit strong cooling flows, feeding massive, ultraluminous central galaxies. Bright galaxies *are* seen at the centres of some rich clusters, but they are generally old ellipticals, not actively star-forming spirals.

It was noticed in the first semi-analytical models that cooling in massive haloes led to an overabundance of bright galaxies, a problem that has led various groups since to come up with different ways of suppressing cooling in the most massive haloes. The

simplest of these is simply to shut off star formation in haloes with circular velocities greater than some empirical value; this was the approach adopted by Kauffmann, White & Guiderdoni (1993), who took $V_c = 500$ km s^{-1} as the largest halo in which stars could form. Somerville (1997) took a similar approach, suppressing cooling completely in haloes with $V_c > 500$ km s^{-1}. Both of these methods had the desired effect of preventing ultraluminous galaxies from forming, but neither provided an explanation for their absence. Somerville & Primack (1999) explained the lack of cooling in massive haloes by means of a 'dynamic halo' model, in which the gas in a newly formed halo was not assumed to have an isothermal profile but, instead, the cooling radius of the largest progenitor was retained. Thus, any new gas was added around the edges of the halo, where the density is lowest and hence the cooling time longest. This is, however, in disagreement with observations of galaxy clusters, which show that the central regions contain large quantities of gas at temperatures close to the virial temperature of the cluster and with densities at which cooling should be efficient. CLBF also suppressed cooling by altering the gas profile, this time by adopting a profile with a constant density core which grows with time as gas cools.

Here, I highlight an alternative solution, proposed by vKJP and illustrated using a simpler version of the model presented here. Explicitly testing the haloes identified in an N-body simulation for virialisation shows that the most massive haloes ar far from being fully virialised. Indeed, they still contain a significant amount of substructure, which can be readily identified in the simulations using local density percolation. This means that, in these haloes, the process whereby gas is shock-heated to the virial temperature of the halo is ongoing. The continuing process of gravitational collapse provides the energy input needed to balance radiative cooling in such haloes. In the model presented here, these massive haloes are separated into their component subhaloes, so that there is no single dominant galaxy in the halo (although each subhalo has its own central galaxy) and the bright end of the LF is determined directly by the cut-off in the subhalo mass function.

Explaining the zero-point of the TFR is a challenge for many galaxy formation models. The fact that the model presented here gives a good fit to both the slope and zero-point of the TFR is promising, but should be treated cautiously since the

assumption of a flat rotation curve for the disc is, at best, an approximation. In reality, the details of the dark matter profile and the self-gravity of the baryons are likely to affect the rotational velocity of the disc. If these effects are important then we can expect the rotational velocities of discs to be significantly higher than the virial velocity of the halo. However, since the observed LF is reproduced well by the model with little room for adjusting the free parameters, it is difficult to see how the observed TFR can be explained unless the rotational velocities of spiral galaxies are close to the virial velocities of their haloes. This, in turn, requires that halo profiles be less centrally-concentrated even than the NFW profile in their central regions. Interestingly, this possibility is also suggested by the rotation curves of low surface-brightness dwarf galaxies (e.g. de Blok & Bosma 2002).

An explanation is also still lacking for the observed numbers of faint galaxies in the universe. This is much less of a problem now than it was a decade ago and significantly less than when the present work was started; as galaxy surveys have probed to deeper isophotal limits, more low surface-brightness galaxies have been uncovered, with no sign of any decline in the population at lower surface brightnesses. This leaves open the possibility that even the current generation of surveys are not complete to their formal magnitude limits at faint absolute magnitudes, because here surface-brightness selection effects start to become important. As galaxy surveys have probed to deeper and deeper isophotal limits, more low-luminosity, low-surface-brightness galaxies have been discovered. It is reasonable to assume, therefore, that the excess faint galaxies predicted by galaxy formation models have even lower surface-brightnesses and are yet to be detected. This was the primary motivation for the work presented in the remaining chapters of this thesis. To account for surface-brightness selection effects, it is neccessary to be able to predict distribution of light in galaxies, which requires, in particular, a model for galaxy discs, since disc galaxies (spirals and lenticulars) make up the largest fraction of the field galaxy population.

2.6 Conclusions

I have argued that a simple phenomenological model, coupled with an N-body simulation with a built-in recipe for the formation and merging of galaxy haloes can be used to model galaxy formation in a cosmological context. The approach used here differs from other models in using halo properties measured directly from the N-body simulation, and in accounting for substructure explicitly. Directly modelling the merging of subhaloes in this way removes a level of uncertainty from the treatment of galaxy–galaxy merging within haloes.

Using the model outlined in sections 2.2 and 2.3, I have shown how overmerging leads to an overabundance of bright galaxies, a problem that can be solved quite simply, by explicitly testing haloes for virialisation, rather than assuming all overdense regions to be virialised. This leaves a population of subhaloes into which gas can cool onto central galaxies whose luminosities match closely those of the most luminous observed galaxies. This match is achieved without the need for any physics more complicated than gravitational collapse, simple radiative cooling, an observationally motivated star formation law and reasonable assumptions about the spectrophotometric properties of stars and dust. I have demonstrated that this result is robust to most reasonable changes in the free parameters of the phenomenological model.

The observed TFR is also well-matched by the model, given the right selection criteria and neglecting the effects of self-gravity.

Chapter 3

Star formation thresholds in galaxy formation models

In this chapter, I build on the basic galaxy formation model set out in chapter 2 by adding a more detailed description of galaxy discs. The main addition is an observationally-motivated star formation law, which includes a surface density threshold. Observationally, it is found that the star formation rate in galaxy discs is a power law of the gas surface density at high densities, but that below some critical density it falls off rapidly. The gas discs of low surface-brightness galaxies (which are known to be gas rich, but have very low star formation rates) are generally found to be below this critical density (e.g. van der Hulst et al. 1993), which suggests that such thresholds play an important role in determining the surface brightnesses of galaxies. I investigate the consequences of star formation thresholds on several properties of disc galaxies, including the distribution of scale lengths as a function of luminosity, which is closely related to the bivariate brightness function. I suggest that star formation thresholds provide an alternative mechanism for achieving some of the effects of feedback, such as suppressing star formation in low circular velocity haloes, by locking up baryons in substantial sub-critical cold gas reservoirs.

3.1 Introduction

The basic model set out in chapter 2 gives a good fit to the observed luminosity functions in B and K, except at the faint end, and reproduces the I-band Tully–Fisher relation. I also showed that these results are not significantly changed by varying the free parameters of the model. In one sense, this is desirable since it allows robust conclusions to be drawn about processes in the dark matter. However, it also means that little can be said about the galaxy formation model itself. Even though I deliberately tried to keep the number of free parameters to a minimum, those that are present in the model (the dynamical friction timescale, star formation efficiency, strength of stellar feedback, to name a few) still allow a large degree of freedom to adjust the model and should prompt the question: *is the solution unique, or are there other models that would fit the data equally well?* This is motivation (if any is needed) to improve the model by adding more physically- or observationally-motivated prescriptions for some of the ingredients of the model and by adding new ingredients, to increase the range of galaxy properties that are predicted. In some places, the basic model is clearly unrealistic; for example, in assuming massless discs and dark matter profiles described by singular isothermal spheres. Making more predictions is one way of more effectively constraining the model. For example, the luminosity function constrains the *number* of galaxies of a given luminosity, but one can imagine two models which both produce the correct number of galaxies at each luminosity, but in which individual galaxies have different luminosities. Using the Tully–Fisher relation as a second constraint helps to break the degeneracy between the models by including another property of the galaxies, namely the circular velocity.

After the total luminosity, the second most important galaxy property is the distribution of the light. To predict this, it is neccessary to have a model for galaxy morphology. The stars in galaxies can be separated into two structural components: a flat, rotating disc and a bulge, or spheroid, which is pressure-supported. In the basic model I already assumed that most star formation occurs in gaseous discs. A common type of star formation law in galaxy formation models has the form

$$\psi_\star = \frac{M_{\rm c}}{\tau_\star}, \tag{3.1}$$

where M_c is the amount of cold gas in the galaxy. The star formation timescale, $\tau_\star$, may be fixed (e.g. Somerville 1997) or it may depend on other properties of the galaxy, such as the circular velocity of the halo (Cole et al. 1994; Baugh, Cole & Frenk 1996; Baugh et al. 1998), the dynamical time of the galaxy (Kauffmann, White & Guiderdoni 1993; Somerville & Primack 1999) or both (CLBF). The common feature of all of these prescriptions is that they assume that the star formation rate is proportional to the amount of cold gas available.

At high surface densities, the azimuthally averaged SFRs of spiral galaxies are well-described by a power-law function of the total (atomic+molecular) gas surface density:

$$\Sigma_{\rm SF} \propto \Sigma_{\rm gas}^n, \tag{3.2}$$

with $n = 1 - 2$ (Schmidt 1959; Kennicutt 1989). This is commonly referred to as a Schmidt law. However, below some critical density, $\Sigma_{\rm crit} \sim 3 - 4$ M$_\odot$ pc^{-2}, the star formation law becomes non-linear and the rate of massive star formation, as indicated by the strength of recombination lines and numbers of HII regions, falls off rapidly with decreasing gas density (Kennicutt 1989). The cut-off in the SFR at low gas densities can be interpreted in terms of the conditions necessary for gravitational instability: at low surface densities, the disc is locally stable to the growth of density perturbations and so molecular clouds, believed to be necessary for star formation, are unable to form. Toomre (1964) showed that a pure gas disc is unstable if

$$Q \equiv \frac{c\kappa}{\pi G \Sigma_{\rm gas}} < 1, \tag{3.3}$$

where c is the velocity dispersion of the gas and κ the epicyclic frequency, given by

$$\kappa = 1.41 \frac{V}{R} \left(1 + \frac{R}{V}\frac{{\rm d}V}{{\rm d}R}\right)^{1/2}. \tag{3.4}$$

Here, V is the rotational velocity of the disc, which will generally be a function of radius R. Kennicutt showed that the Toomre condition (equation 3.3) can be written in the form $\Sigma_{\rm gas} > \Sigma_{\rm crit}$, where

$$\Sigma_{\rm crit} = \alpha_{\rm K} \frac{\kappa c}{3.36 G}. \tag{3.5}$$

In regions where $\Sigma_{\rm gas} < \Sigma_{\rm crit}$ the disc is stable and molecular cloud complexes, necessary for the formation of large numbers of massive stars, do not readily form. The

parameter α_K is then chosen to match the critical radii of galaxies (the radii at which the azimuthally averaged gas density falls below the critical density in each case) to the observed cut-offs in the distributions of HII regions, which are associated with young, massive stars. The existence of star formation thresholds affects the evolution of galaxies in two ways: by limiting the amount of gas that can be turned into stars and by defining the star-forming region of the disc, thus determining the amount of gas considered 'available' for star formation (Martin & Kennicutt 2001). Some galaxies, mainly those of low surface brightness, have gas discs that are entirely below this critical density.

The sizes of galaxy discs are fundamentally determined by the balance between gravity and the angular momentum of the disc material. The gravitational force has components not only from the dark matter, as assumed in chapter 2, but also from the baryons themselves. The contribution from the latter can be important in the cores of haloes. There is also the possibility that the halo itself is not well described by the singular isothermal sphere. In fact, Navarro, Frenk & White (1997, NFW) found that the haloes identified in N-body simulations are better fit by a universal density profile (equation B.5), which has a shallower slope than the isothermal sphere in the central regions and is more extended in the outer regions. Hydrodynamic simulations (e.g. Navarro, Frenk & White 1995) suggest that a significant fraction of the initial angular momentum of the baryons is lost during galaxy formation, but this is inconsistent with observations of spiral galaxies, producing discs which are too compact.

In this chapter I present a model for galaxy morphology which includes the effects of star formation thresholds, self-gravity and 'realistic' halo profiles on the properties of discs. Predicting the sizes of discs gives a stronger constraint on the galaxy formation model than the luminosity function alone. It is also important in quantifying the effects of surface brightness limits in observational samples, since it is the distribution of light, and not the total flux, that determines whether or not a galaxy will be included in a sample.

3.2 Model

3.2.1 Overview of the basic ingredients

The work described in this chapter builds on the basic model presented in chapter 2. Here I present an overview of the main ingredients of this model (this overview was adapted from vKJP).

(1) *The merging history of dark matter haloes.* This is taken directly from an N-body simulation. Special techniques prevent galaxy-scale haloes undergoing 'over-merging' owing to inadequate numerical resolution. By allowing substructure to survive, the merger history is simplified and the number of halo mergers is reduced.

(2) *The merging of galaxies within dark matter haloes.* Each halo contains a single galaxy at formation. When haloes merge, a criterion based on dynamical friction is used to decide how many galaxies exist in the newly formed halo. The most massive of those galaxies becomes the single central galaxy onto which gas can cool, while the others become its satellites.

(3) *The thermal history of gas within dark matter haloes.* A halo forms with a mass $M_{\rm h}$ of hot gas (for a new halo $M_{\rm h} = (\Omega_{\rm b}/\Omega_{\rm M})M_{\rm halo}$, but in general the hot gas reservoir will have been depleted by cooling and star formation at earlier times), which is assumed to quickly settle to the virial temperature of the halo. This gas can cool to form stars within a single galaxy at the centre of the halo; application of the standard radiative cooling curve (Sutherland & Dopita 1993) shows the rate at which this hot gas cools below 10^4 K, and is able to form stars. When haloes merge, all hot gas is stripped from the individual galaxies and is added to the hot gas reservoir of the new halo, from where it can cool onto the central galaxy. Thus, each halo maintains an internal account of the amounts of gas being transferred between the two phases, and consumed by the formation of stars.

(4) *Feedback from star formation.* Energy output from supernovae reheats some of the cooled gas back to the hot phase. This ingredient is essential to keep low-mass galaxies and proto-galaxies from using up all their gas at high redshift.

(5) *Chemical evolution.* Evolving stellar populations inject gas and metals, as well as energy, into the interstellar medium. The evolution of the metals must be followed, because the cooling of the hot gas depends on metal content, and a stellar population of high metallicity will be much redder than a low metallicity one of the same age. I adopt the instantaneous recycling approximation and assume that all of the metals are initially ejected into the cold gas.

(6) *Star formation in discs.* Stars form according to an observationally-motivated star formation law, in discs whose sizes are determined by the angular momentum of the baryons. A fraction $\mathcal{R}$ of the gas turned into stars is returned to the ISM by stellar winds and supernovae, thus increasing the gas consumption timescale.

(7) *Starbursts.* The SFR may suffer a sharp spike following a major merger event. This is motivated empirically by the existence of ultra-luminous IRAS galaxies, and it allows hierarchical models to yield behaviour resembling traditional monolithic collapse.

(8) *Stellar spectral synthesis.* Having formed stars, we wish to predict the appearance of the galaxy that results. For this, it is necessary to assume an IMF, and to have a spectral synthesis code. Here, I use the spectral models of Bruzual & Charlot (1993), which allow synthetic stellar populations of any metallicity to be constructed. The IMF is taken to be Salpeter, but this is only one of several choices available. I use a simple scaling law prescription, based on observations, to correct for the effects of dust absorption on the spectra of galaxies.

3.2.2 Star formation in discs with thresholds

I assume the gas surface density to be exponential, as before. If the rotation curve of the disc is approximately flat then equation (3.4) takes the simple form $\kappa = 1.41 V_c/R$. This is, of course, slightly unrealistic: if self-gravity is important, which it must be for gravitational stability to be important, then an entirely flat rotation curve is impossible. However, the observed rotation curves of spiral galaxies are flat at large radii, so the approximation cannot be too far from the truth. I follow Kennicutt in assuming a value of $c = 6$ km s^{-1} for the velocity dispersion in the disc. The critical density in this case

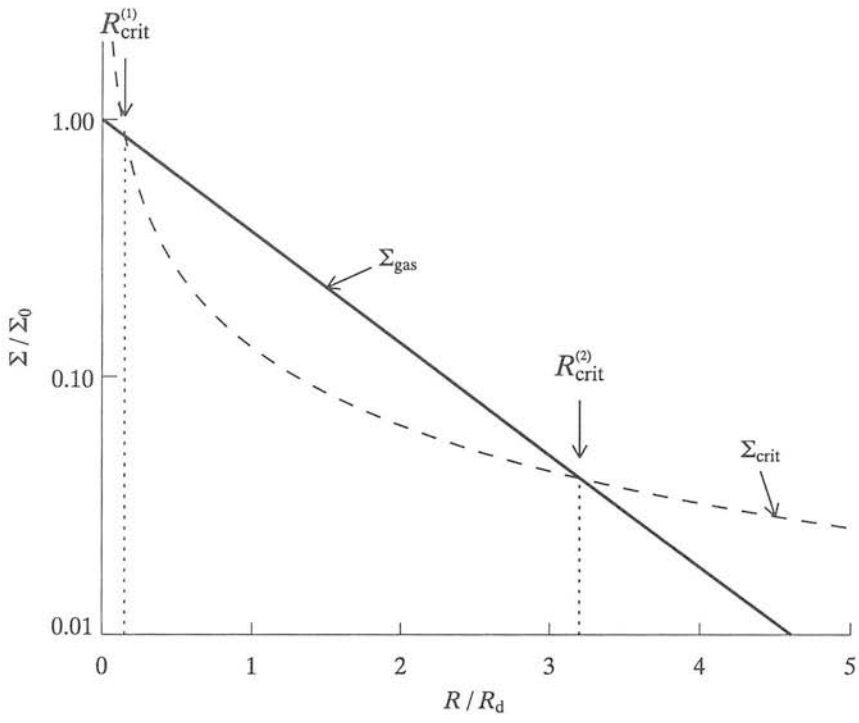

Figure 3.1: Kennicutt's star formation threshold for a galaxy with a flat rotation curve and an exponential surface density profile. The gas density (solid line) is above the critical density for star formation (dashed line) between the two critical radii, $R^{(1)}_{\rm crit}$ and $R^{(2)}_{\rm crit}$.

is

$$\Sigma_{\rm crit}(R) = 0.59\alpha_{\rm K}\frac{V_{\rm c}/\ {\rm km\,s^{-1}}}{R/\ {\rm kpc}}\ {\rm M_\odot pc^{-2}}. \tag{3.6}$$

A value of $\alpha_{\rm K} = 0.74 \pm 0.2$ then gives a good fit to observed threshold radii (Martin & Kennicutt 2001) in spiral galaxies. This form of the critical density has the qualitative properties that it decreases more steeply with radius than an exponential at small radii but has a shallower radial dependence at large radii. Hence, if the gas density rises above the critical density at all, there will be a region, bounded by two critical radii, $R^{(1)}_{\rm crit}$ and $R^{(2)}_{\rm crit}$, where the gas density is above the critical density and stars can form. Fig. 3.1 illustrates the situation in a typical galaxy. I identify the region between $R^{(1)}_{\rm crit}$ and $R^{(2)}_{\rm crit}$ with the optical disc of the galaxy.

I adopt a model in which star formation only occurs in the region between $R^{(1)}_{\rm crit}$ and $R^{(2)}_{\rm crit}$, where the gas surface density is above the critical density. Stars form according

to the same law as before:

$$\psi_{\rm SF} = \frac{M_{\rm avail}}{\tau_\star}, \tag{3.7}$$

except that the mass of available gas, $M_{\rm avail}$, is now taken to be the mass of gas within the star forming region, which I identify with the optical disc of the galaxy. Recall that $\tau_{\rm disc}$ is defined to be the orbital time at the half-mass radius of the disc:

$$\tau_{\rm disc} \equiv 2\pi R_{\rm half}/V_{\rm c}. \tag{3.8}$$

In the absence of star formation, the mass of gas 'available' for star formation at time t would be

$$M^0_{\rm avail} = M_{\rm cooled}\left[\left(1+\frac{R^{(1)}_{\rm crit}}{R_{\rm d}}\right)e^{-R^{(1)}_{\rm crit}/R_{\rm d}} - \left(1+\frac{R^{(2)}_{\rm crit}}{R_{\rm d}}\right)e^{-R^{(2)}_{\rm crit}/R_{\rm d}}\right], \tag{3.9}$$

where $M_{\rm cooled}$ is the mass of hot gas that has cooled since the halo formed. Star formation and reheating at earlier times will have depleted the reservoir of available gas, so the actual mass available is

$$M_{\rm avail} = M^0_{\rm avail} - M_\star - M_{\rm rh}, \tag{3.10}$$

where $M_\star = (1-\mathcal{R})M_{\rm SF}$ is the mass of long-lived stars and $M_{\rm rh}$ the mass of gas returned to the hot phase by supernovae. Star formation can only continue until the surface density reaches the threshold given by equation (3.5) so, in practice, only some fraction of the gas in the star-forming region is able to form stars. The fraction of the gas between the two critical radii which lies below the threshold and is thus unable to be turned into stars is

$$\begin{aligned} f_{\rm sub} &= \frac{2\pi}{M_{\rm avail}}\int_{R^{(1)}_{\rm crit}}^{R^{(2)}_{\rm crit}} \Sigma_{\rm crit}(R)\,R\,dR \\ &= 3707\left(\frac{V_{\rm c}}{\rm km\ s^{-1}}\right)\left(\frac{R^{(2)}_{\rm crit}-R^{(1)}_{\rm crit}}{\rm pc}\right)\Big/\left(\frac{M_{\rm avail}}{\rm M_\odot}\right). \end{aligned} \tag{3.11}$$

Equation (2.39) thus becomes

$$\begin{aligned} M_{\rm SF} &= \frac{M^0_{\rm avail}}{1+\beta-\mathcal{R}} \times \\ &\quad \min\left\{\left[1-e^{-(1+\beta-\mathcal{R})\epsilon_\star\frac{\Delta t}{\tau_{\rm dyn}}}\right], 1-f_{\rm sub}\right\}. \end{aligned} \tag{3.12}$$

So the existence of a threshold affects the star formation law in two ways. Firstly, it restricts the range of radii over which star formation can take place and hence reduces the mass of gas that is considered 'available' for star formation and, secondly, it sets an upper limit to the fraction of this gas that can actually form stars.

3.2.3 Disc scale lengths and rotation velocities

The rotation curve of a halo with an NFW profile is given by equation (B.7). Mo, Mao & White (1998, MMW) give a useful approximation to the scale length of a self-gravitating disc in a halo with an NFW profile (MMW expressed the following in terms of λ and r_{200}; here I have used equation 2.36 to express the scale length of a non-self-gravitating disc in an isothermal sphere halo in terms of the more natural quantities $j_{\rm disc}$ and $V_{\rm c}$):

$$R_{\rm d} = f_R\, f_c^{-1/2}\, \frac{j_{\rm disc}}{2V_{\rm c}}, \tag{3.13}$$

where the factor f_c (equation B.9) gives the effect of changing the binding energy of the halo, while the factor

$$f_R \simeq \left(\frac{\lambda'}{0.1}\right)^{-0.06+2.71m_{\rm d}+0.0047/\lambda'} (1-3m_{\rm d}+5.2m_{\rm d}^2) \times(1-0.019c+0.00025c^2+0.52/c) \tag{3.14}$$

gives the remaining dependence on the density profile. Here, $m_{\rm d} = M_{\rm disc}/M_{\rm halo}$ is the total mass (gas+stars) of the disc, expressed as a fraction of the halo mass, and $\lambda' = f_j\lambda$, where λ is the spin parameter of the halo and the parameter f_j describes the fraction of the initial angular momentum retained during the disc formation process, such that $j_{\rm disc} = f_j j_{\rm halo}$.

MMW also give an expression for the maximum rotational velocity of the disc (in practice, the velocity measured at $3R_{\rm d}$, close to the peak of the rotation curve):

$$V_{\rm max} = f_V\, V_{\rm c}, \tag{3.15}$$

where

$$f_V \simeq \left(\frac{\lambda'}{0.1}\right)^{-2.67m_{\rm d}-0.0038/\lambda'+0.2\lambda'} (1+4.35m_{\rm d}-3.76m_{\rm d}^2) \times\frac{1+0.057c-0.00034c^2-1.54/c}{[-c/(1+c)+\ln(1+c)]^{1/2}}. \tag{3.16}$$

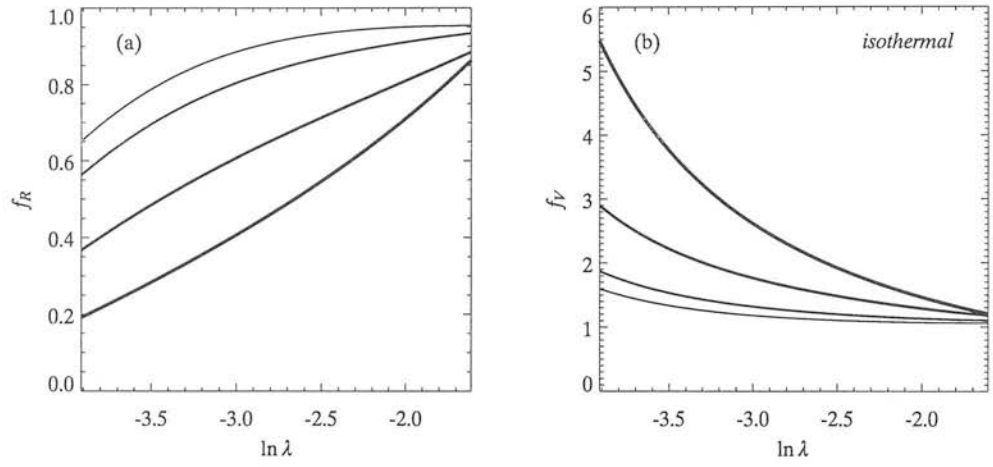

Figure 3.2: Approximate corrections to (a) the disc scale length and (b) the maximum rotational velocity of the disc for the effects of self-gravity. See text for definitions and details. The correction factors are plotted as functions of λ, the spin parameter of the halo, assuming $f_j = 1$. The curves in each panel are for (lightest to heaviest) $m_d = 0.02$, 0.04, 0.1 and 0.2.

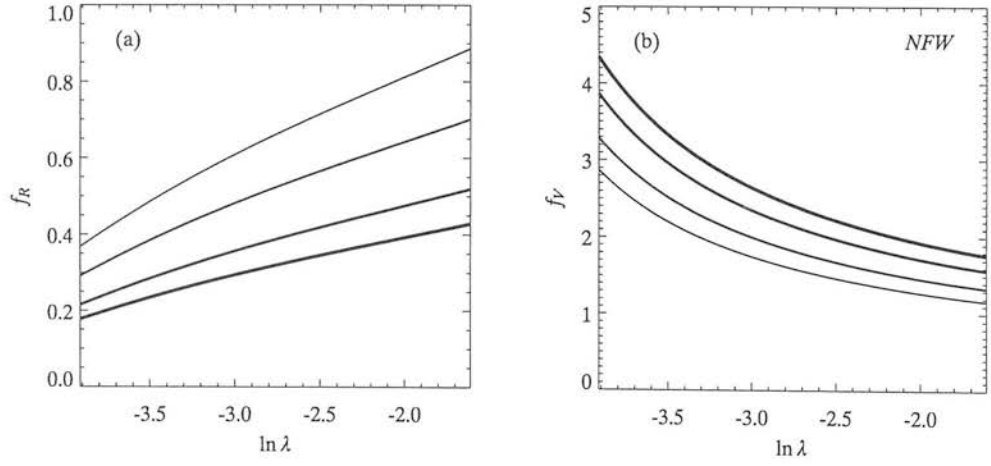

Figure 3.3: As Fig. 3.2 for the combined effects of self-gravity and the NFW halo profile. See text for definitions and details. The curves in each panel are for concentration parameters of (lightest to heaviest) $c = 5$, 10, 20 and 30.

In Fig. 3.2 I plot the correction factors to the scale length and circular velocity from disc self-gravity only (i.e. setting $f_c = 1$ and ignoring the terms in equations 3.14 and 3.16 involving only c), as functions of λ, for different disc masses. I assume $f_j = 1$; for other values it is simply neccessary to exchange λ for λ' on the horizontal axis. In Fig. 3.3 I plot the corrections for different values of the NFW concentration parameter, c[1], assuming a fixed disc mass of $m_\mathrm{d} = 0.1$. In all cases the corrections are in the sense of higher rotational velocities and smaller scale lengths, with the lowest angular momentum discs being the most affected.

The combined effects of self-gravity and the different halo profile means that, in general, disc rotation curves will no longer be flat. The shape of the rotation curve affects the form of the star formation threshold through the epicyclic frequency (equation 3.4). To separate this effect from the overall effect of changing the sizes of discs, I retain, for now, the assumption of a flat rotation curve when calculating the critical density and I substitute the maximum rotation velocity, V_max, for V_c in equation (3.6).

Equations (3.14) and (3.16) are said to be accurate to within 15% for the ranges $5 < c < 30$, $0.02 < \lambda' < 0.2$ and $0.02 < m_\mathrm{d} < 0.2$. Where model galaxies fall outside these ranges for one of the parameters, which happens occasionally, I simply evaluate the equations assuming the extreme value of the parameter in question.

A brief note is neccessary here on the way that the halo properties are calculated for different profiles. The MMW approach has as its basis the spherical collapse model, which predicts the densities of haloes. Changing the density profile thus has the effect of changing the binding energy of the halo (by a factor f_c for the change from an isothermal sphere to the NFW profile). My approach is to measure the binding energy of haloes from the N-body simulations. This is fixed and it is the mean overdensity, and hence the virial radius and velocity, that changes, in the latter case by a factor $f_c^{-1/2}$. This means that the factor $f_c^{-1/2}$ should not be present in equation (3.13) for the disc scale length. Unfortunately, no such simple fix exists for the rotational velocity (equation 3.15), so for now I take the MMW approximations at face-value.

[1] To calculate c, I use the method of Peacock & Smith (2000), which differs from that originally employed by NFW.

3.2.4 Forming bulges and ellipticals

An important part of any model for disc galaxies is the formation of bulges (and ellipticals). This is particularly true of models in which the majority of stars are assumed to have formed in discs and and were subsequently transferred to a spheroidal component. There are several possible mechanisms for bulge formation. The three most important are: (i) collisionless mixing of stellar orbits during collapse; (ii) mergers between disc galaxies; and (iii) secular evolution of the disc, possibly involving bar-like stabilities. I consider only the second of these: the formation of spheroids in galaxy–galaxy mergers. This is the most likely mechanism for the formation of large bulges and ellipticals, while secular evolution may play a role in creating small bulges in late-type galaxies.

When two disc galaxies merge, the morphology of the remnant depends primarily on the mass ratio of the two galaxies involved (it also depends, to a lesser extent, on the orbital parameters and the relative orientations of the discs). Several studies (Barnes 1992; Hernquist 1992) using N-body simulations have shown that mergers between equal-mass disc galaxies produce elliptical remnants with $r^{1/4}$–law light profiles; the discs of both galaxies are completely destroyed. In contrast, Walker, Mihos & Hernquist (1996) have shown that the accretion by a disc galaxy of a satellite with mass 1/10th that of the accreting galaxy results in a thickening of the disc, but does not destroy it. In this case, the core of the satellite falls to the centre of the more massive galaxy. Recent studies (Naab, Burkert & Hernquist 1999; Naab & Burkert 2001a) suggest that the critical mass ratio dividing these two regimes is around 4 : 1, with 3 : 1 mergers producing ellipticals which have disc-like isophotes. In both of the cases outlined above, the merger results in an inflow of gas, triggering a period of enhanced star formation activity (Mihos & Hernquist 1994a, b, 1996). This burst of star formation is particularly strong in major mergers, where it may start as early as the time of the first close approach of the galaxies and continues until they finally merge.

The picture is thus one of constantly evolving galaxy morphologies, as illustrated in Fig. 3.4. Initially, gas cools onto a disc and forms stars. If such a disc is involved in a merger with another of similar mass then both discs are destroyed and a spheroidal

merger remnant is produced. Subsequently, fresh gas can cool around the remnant, forming in a new disc, in which a new generation of stars can form, and the process is repeated. Alternatively, a disc galaxy can grow a bulge by accreting many small satellites, without its own disc being disrupted.

I define the mass ratio, $M_{\rm sat}/M_{\rm cen}$, of two merging galaxies to be the ratio of the total baryonic masses (gas+stars) of the discs of the two galaxies. I define two regimes:

$$\begin{aligned} M_{\rm sat}/M_{\rm cen} &\geqslant f_{\rm b} \quad \textit{major merger}, \\ M_{\rm sat}/M_{\rm cen} &< f_{\rm b} \quad \textit{minor merger}, \end{aligned} \tag{3.17}$$

separated by a critical mass ratio, $f_{\rm b}$, which is a free parameter of the model.

(1) *Major mergers.* When two galaxies of mass ratio $M_{\rm sat}/M_{\rm cen} \geqslant f_{\rm b}$ merge, all of the stars in the two galaxies are transferred to the bulge of the merger remnant.

(2) *Minor mergers.* When the central galaxy in a halo accretes a satellite with $M_{\rm sat}/M_{\rm cen} < f_{\rm b}$, all of the stars in the satellite (both disc and bulge) are transferred to the bulge of the central galaxy. The central galaxy retains its own disc.

In the event of a major merger, any cold gas is turned to stars in a single burst of star formation; these stars are placed in the bulge of the merger remnant. No starburst results from a minor merger. Note that these starbursts are different, and in addition to, those caused by halo mergers. When a galaxy undergoes several merger events in a single timestep, I treat the events separately, so that the disc is destroyed if any one of the mergers satisfies $M_{\rm sat}/M_{\rm cen} \geqslant f_{\rm b}$. I take no account of the cumulative effect on a galaxy of consecutive, closely-spaced mergers, which might be expected to have a greater effect than each taken on its own.

Bursts on galaxy–galaxy mergers make a small contribution ($\sim 5\%$) to the present-day stellar content of the universe. Fig. 3.5 shows the star formation history of the universe, broken down into contributions from different modes of star formation, for a model with bursts of star formation on major mergers. At the present day, such bursts make a contribution to the star formation rate comparable to that from starbursts driven by halo mergers, but their relative importance falls off more quickly with redshift.

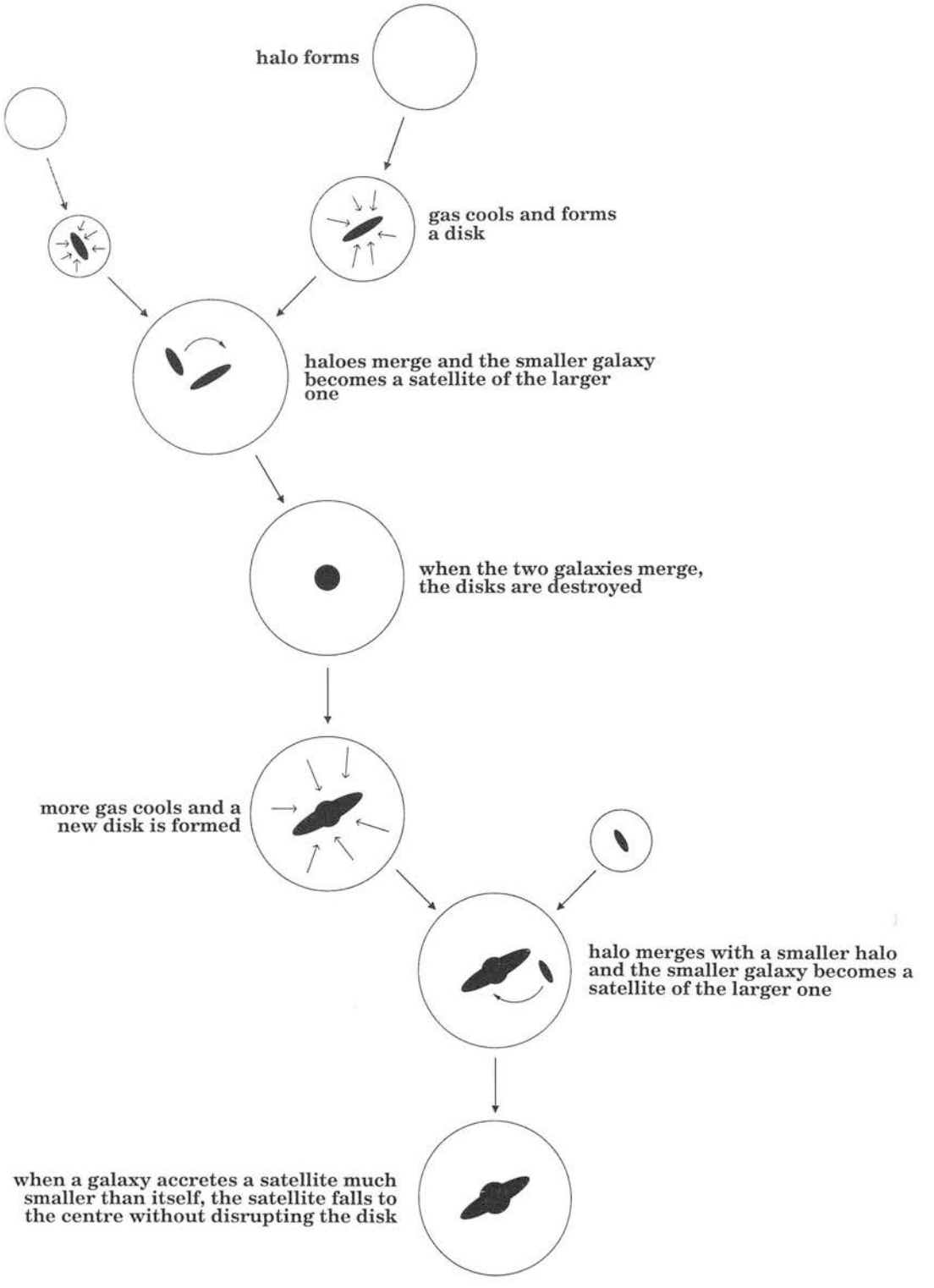

Figure 3.4: A merger-driven model for the evolution of galaxy morphologies. As gas cools and collapses in haloes it forms discs, but these can be destroyed in major mergers, when two galaxies of comparable mass merge, resulting in a spheroidal merger remnant. Further cooling of gas onto the remnant can subsequently form a new disc, with the spheroidal component as the central bulge.

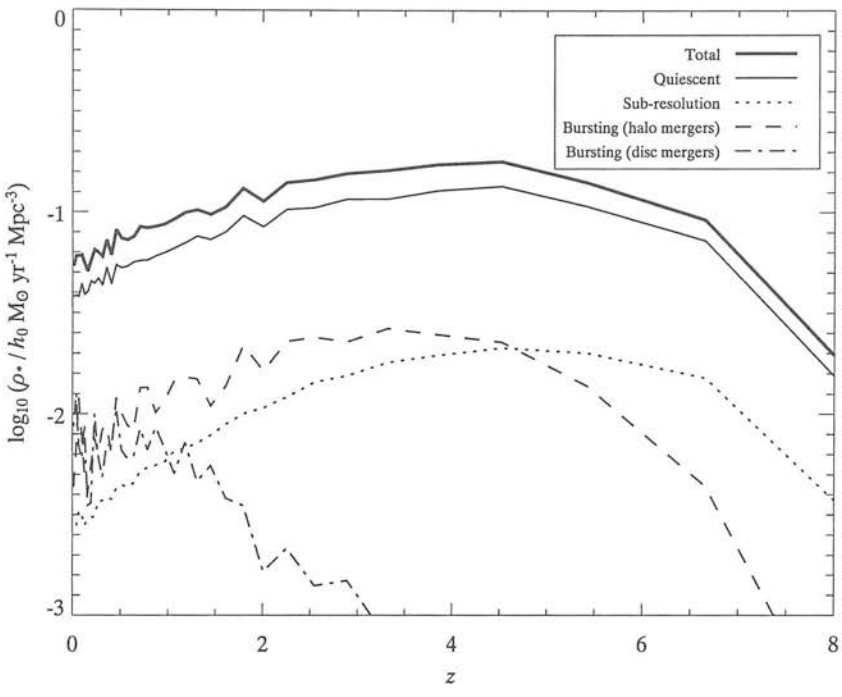

Figure 3.5: The star formation history of the universe as a function of redshift, including a contribution from bursts on galaxy–galaxy mergers (dot-dashed line). Curves are plotted for model K (see table 3.1). Contributions from other modes of star formation are also shown: quiescent star formation in discs (light, solid line), starbursts (dashed line) and star formation in sub-resolution haloes (dotted line).

Dusty spheroids

In the basic model I assumed that the spheroidal component of galaxies was unaffected by dust extinction. This is generally true for old bulges and ellipticals. In the above model, spheroids form in major mergers, often with an associated burst of star formation. These young stars are placed in the spheroid and, because of the lack of dust extinction, are extremely bright. Of course, this is just an indication that the model is too simple in this respect — galaxies that have undergone a recent major merger will not have settled down into their final spheroidal configurations and will have chaotic morphologies. Such recent mergers are observed as ultraluminous infrared galaxies (ULIRGs) and are found to contain significant amounts of dust, associated with the starburst.

Table 3.1: Parameters of the disc models.

	Model						
Parameter	basic	A	B	K	G	N	Eqn.
$f_{\rm b}$		0.25	0.25	0.25	0.25	0.25	(3.17)
$\alpha_{\rm K}$				0.7	0.7	0.7	(3.6)
$f_{\rm mrg}$	1.0	1.0	0.5	0.5	0.5	0.5	(2.20)
$\epsilon_\star$	0.05	0.05	0.05	0.05	0.005	0.005	(2.33)
self-gravity					•	•	
NFW profile						•	

The starburst phase is not modelled in detail, so I adopt a simple model in which the spheroids of galaxies that have undergone a merger in the last Gyr are assumed to suffer from dust extinction. For simplicity I adopt the same dust extinction law as for discs (equation 2.45), which at least ensures that the bright end of the luminosity function is unchanged by changes in the morphologies of galaxies.

3.3 Properties of disc galaxies

The simplest way to illustrate the effects of the several ingredients of the disc model is to start from the basic model set out in chapter 2 and to then add the new ingredients one by one. In table 3.1 I define five new models, in addition to the basic model. The parameters of the basic model are unchanged in the other models, with the exception of the star formation efficiency, $\epsilon_\star$, which has to be reduced in models with self-gravity, to allow for the smaller dynamical times.

3.3.1 Bulge-to-disc ratios

Since, in this section, I am interested in the properties of disc galaxies, it is important to first define what is meant by this term. The meaning may seem obvious but, in fact, there are very few galaxies that are pure discs (mainly LSB galaxies and very late-type

spirals). The Hubble sequence is a continuum in bulge-to-disc ratio with pure discs at one end of the spectrum and at the other pure spheroids (ellipticals). Between these lie bulge-dominated spirals, lenticulars, and ellipticals with discy isophotes.

The bulge-to-disc ratio is a convenient variable to use for morphological classification in galaxy formation models (e.g. Kauffmann, White & Guiderdoni 1993; Somerville & Primack 1999; CLBF). It correlates strongly with morphologies assigned by visual inspection of images (Simien & de Vaucouleurs 1986), although there is a significant overlap, particularly at the important dividing line between spirals and lenticulars. This is not surprising, since the bulge-to-disc ratio is only one of several criteria for classifying disc galaxies in the Hubble scheme. The distinction between lenticulars and spirals aside, the relative importance of the two main structural components is probably the most physically significant way of classifying galaxy morphologies. For the purposes of this section, I divide galaxies into three classes, on the basis of their bulge fractions (the fraction of the total flux being emitted by the spheroidal component) in the B band:

$$\begin{aligned} (\mathrm{B/T})_B > 0.6 \quad & \text{bulge-dominated galaxies,} \\ 0.4 < (\mathrm{B/T})_B \leqslant 0.6 \quad & \text{intermediate galaxies,} \\ (\mathrm{B/T})_B \leqslant 0.4 \quad & \text{disc-dominated galaxies.} \end{aligned}$$

The boundaries correspond approximately to the divisions between ellipticals, lenticulars and spirals according to the results of Simien & de Vaucouleurs, and it is tempting to identify the morphological classes defined above with those of the Hubble scheme. However, this should be done with caution since the overlap between the bulge-to-disc ratios of the different Hubble types will, at the very least, increase the scatter in the predicted properties of any individual class.

The main parameters affecting the bulge fractions of galaxies are the merging timescale, f_{mrg} and the critical mass ratio for a major merger, f_{b}. In Fig. 3.6 I plot the distribution over bulge fraction, $(\mathrm{B/T})_B$, for the galaxies in model A, divided into different ranges into absolute magnitude. The different classes of galaxy defined above are marked by shading of the histogram: white for disc-dominated galaxies, dark grey for bulge-dominated and paler grey for intermediate galaxies. The dotted histogram shows the same result for the basic model of chapter 2; the difference between these

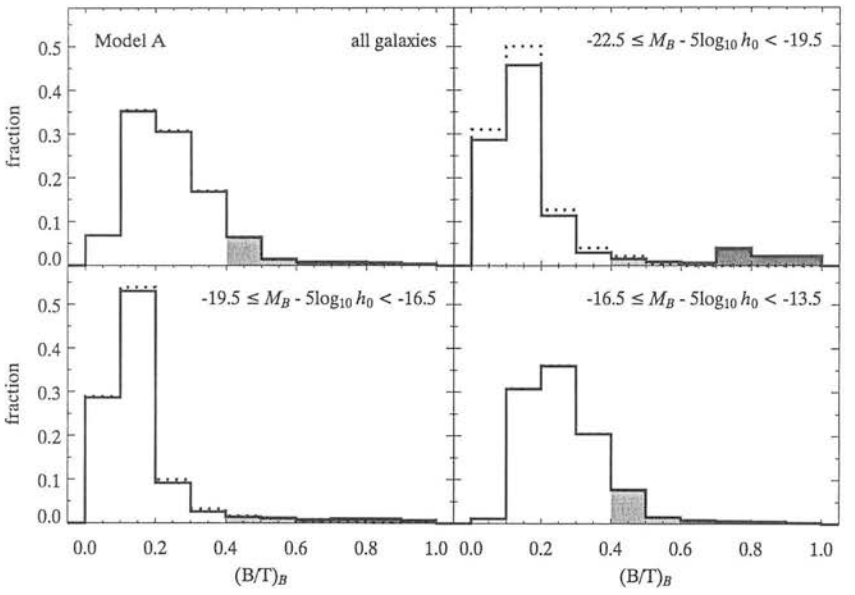

Figure 3.6: The bulge fractions, $(B/T)_B$, of galaxies in the B band for model A, showing the effect of changing f_b. The top-left panel shows the distribution for all galaxies, while in the remaining three panels the galaxies are divided into absolute magnitude bins. Model A is shown by the solid histogram, with shading corresponding to the morphlogical classes defined in the text. Galaxies in the white bins are disc-dominated, while those in darkest bins are bulge-dominated and the bins containing galaxies with intermediate properties are a lighter shade grey. The dotted histogram is for the basic model.

two models illustrates the effect of distinguishing between major and minor mergers. The ratios of the number of galaxies in each morphological class for model A is *disc* : *intermediate* : *bulge* = 89 : 2 : 9. If these classes are identified with the Hubble types E, S0 and S+Irr then the distribution is biased towards the two extremes of almost no disc and almost no bulge, compared to the morphological mix in the APM Bright Galaxy Catalogue (Loveday 1996): S+Irr : S0 : E = 67 : 20 : 13. Most importantly, from the point of view of the remainder of this section, model A predicts too many disc-dominated, or spiral, galaxies. In Fig. 3.7 I plot the distributions for model B (again, the dotted histogram is for the basic model). For this model, the morphological ratios in the brightest bin are *disc* : *intermediate* : *bulge* = 77 : 5 : 18, which is closer to the observed ratios, although still rather biased towards the extreme ends of the spectrum. Adding a star formation threshold (model K, Fig. 3.8) reduces the number of disc-

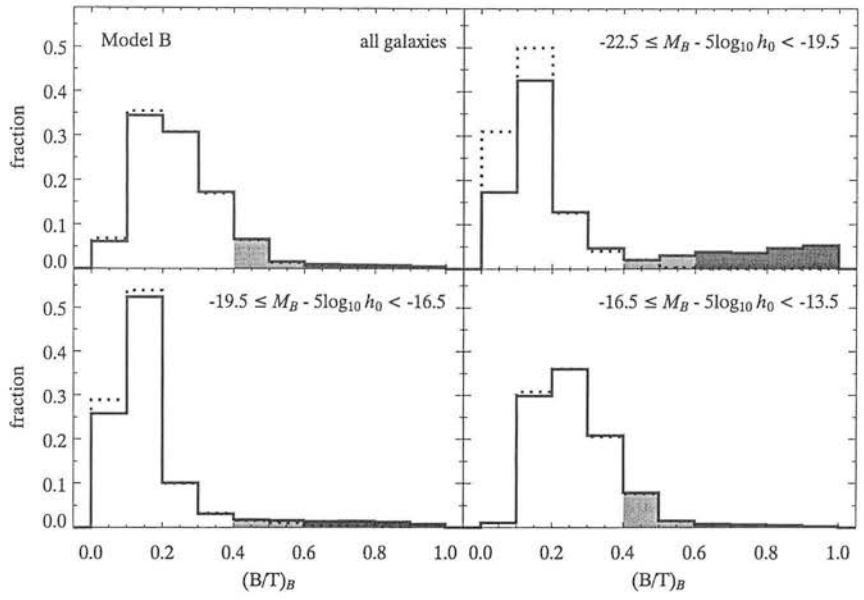

Figure 3.7: As Fig. 3.6 for model B, showing the effect on the bulge fractions of changing $f_{\rm mrg}$.

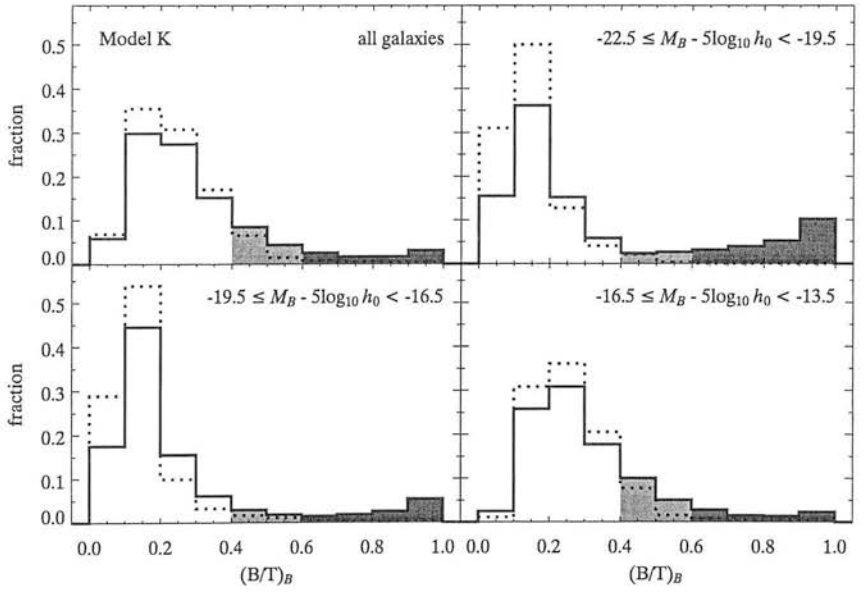

Figure 3.8: As Fig. 3.6 for model K, showing the effect on the bulge fractions of adding a star formation threshold.

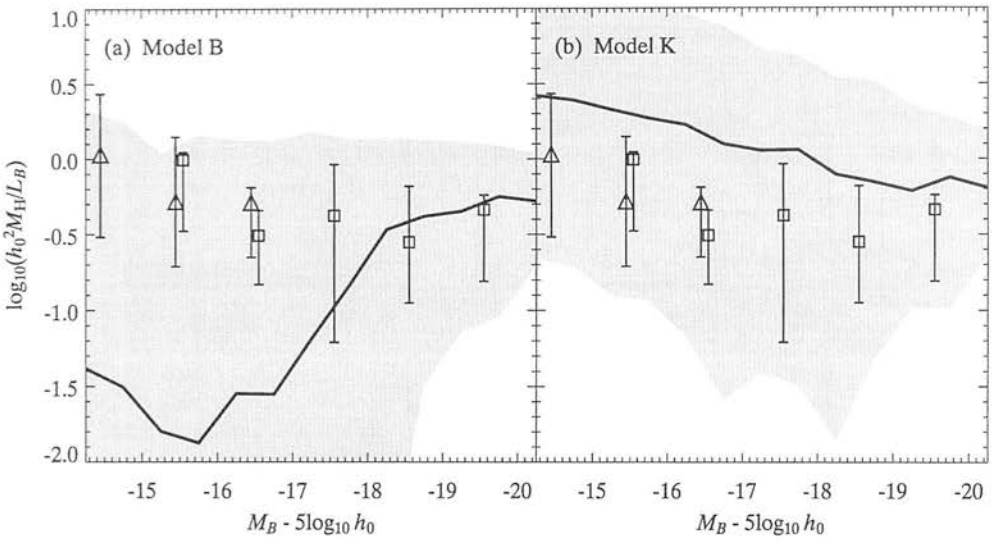

Figure 3.9: Gas contents of spiral galaxies as a function of B-band magnitude. In (a) I plot the distribution for model B, which has no star formation threshold. The heavy, solid line is the median gas mass-to-light ratio for all galaxies with $(\mathrm{B/T})_B < 0.4$ and the shaded region indicates the 10% and 90% quartiles of the distribution. In (b) I plot the distribution for model K. Symbols with error bars are observational measurements as follows: open squares are HI only measurements for a complete sample of spiral galaxies of all Hubble types (Huchtmeier & Richter 1988); open triangles are HI+H_2 measurements for a sample of Sa-Sd galaxies (Sage 1993). In both cases the data points and error bars represent the median and 10 and 90% quartiles of the observed distributions.

dominated galaxies slightly, at the expense of bulge-dominated ones (72 : 5 : 23), but the number of intermediate galaxies remains the same. It appears to be quite difficult to produce galaxies with bulge-to-disc ratios of order unity in this model, although it should be noted that bulge-to-disc ratio is not the cleanest way to separate spirals and lenticulars. Indeed, it forms no part of the observational criteria for distinguishing these two classes.

3.3.2 Gas contents of spiral galaxies

The property of galaxies most strongly affected by the presence of a star formation threshold is the mass of cold gas in the disc. As shown in section 3.2.2, the model gives definite predictions for the fraction of the cold gas that is allowed to be turned into stars (although the threshold does not prevent some, or all, of the remaining gas

being reheated to the hot phase as a consequence of star formation). Observationally, the most convenient measure of the gas fractions of galaxies is the ratio of the gas mass to the luminosity in some band. In Fig. 3.9 I plot the distribution of hydrogen mass-to-light ratio in the B band, as a function of absolute B-band magnitude for disc-dominated galaxies in models B and K, which differ only by the presence of the star formation threshold in model K. The mass of hydrogen in the model galaxies is calculated assuming a 77 : 23 mix of H : He and includes both atomic (HI) and molecular (H_2) hydrogen phases. The observations plotted in the figure are from two sources: for low-luminosity, late-type galaxies, the molecular fraction is generally low (~15%; Boselli, Lequeux & Gavazzi 2002), so the Huchtmeier & Richter points should give a good approximation to the total Hydrogen mass, while at bright magnitudes the field population is dominated by early-type spirals, so the Sage data should be a good estimate of the overall distribution for spirals of all types.

For the brightest galaxies in model B, the predicted gas masses are a fair match to the observational estimates. The model, however, predicts a correlation between luminosity and the median gas mass-to-light ratio in the opposite sense to what is observed and with a scatter orders of magnitude larger at faint magnitudes. Contrast this with the distribution for galaxies in model K. The star formation threshold mainly affects the faintest galaxies: the gas contents of these galaxies are almost entirely determined by the threshold, which locks up baryons in the cold gas reservoir and prevents them from being turned into stars. Model K predicts a correlation of in the correct sense and slope, although the median gas fractions are systematically high and the scatter too large.

The discrepancy between the gas contents of the model galaxies and those in the observational samples may, at first sight, seem to be a problem for the model. However, it is likely that the observational samples are biased towards high surface-brightness (HSB) galaxies. LSB galaxies are generally observed to have higher gas fractions than their HSB counterparts, so if a significant fraction of the model galaxies have low surface brightnesses then their presence will bias the average gas content towards higher values. In Fig. 3.10 I plot the Hydrogen mass-to-light ratio as a function of disc central surface brightness. From this plot it is clear that the galaxies with the highest gas fractions

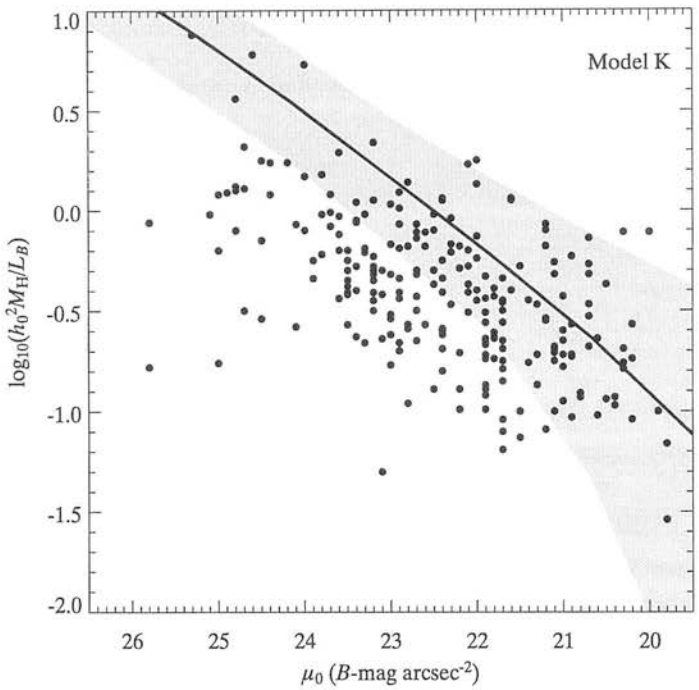

Figure 3.10: Gas contents of spiral galaxies as a function of disc central surface brightness. The heavy, solid line is the median gas mass-to-light ratio for all galaxies with $(B/T)_B < 0.4$ and the shaded region indicates the 10% and 90% quartiles of the distribution. Filled circles are observational measurements of HI mass-to-light ratios by Burkholder, Impey & Sprayberry (2001). For high surface-brightness galaxies (those with $\mu_0 \leqslant 23$ $B-$mag arcsec^{-2}) I correct for the molecular contribution using the results of Young & Knezek (1989), while for low surface-brightness galaxies ($\mu_0 > 23$ $B-$mag arcsec^{-2}) I plot the raw HI measurements.

are indeed of low surface-brightness, such that they would not be included in even the deepest observational samples. Burkholder, Impey & Sprayberry (2001) have measured the HI contents of a sample of HSB and LSB galaxies and I plot their data in Fig. 3.10. HSB galaxies are expected to have a significant fraction of molecular Hydrogen; Young & Knezek (1989) have measured M_{H_2}/M_{HI} as a function of Hubble type, which I use to apply an average correction to the observations of Burkholder et al. for galaxies with $\mu_0 \leqslant 23$ $B-$mag arcsec^{-2} (the value of 23 $B-$mag arcsec^{-2} is arbitrary, but is often taken as an approximate division between HSB and LSB galaxies). For galaxies with $\mu_0 > 23$ $B-$mag arcsec^{-2} I plot the raw HI measurements, since LSB galaxies are expected to have much lower molecular fractions (e.g. Schombert et al. 1990). The gas

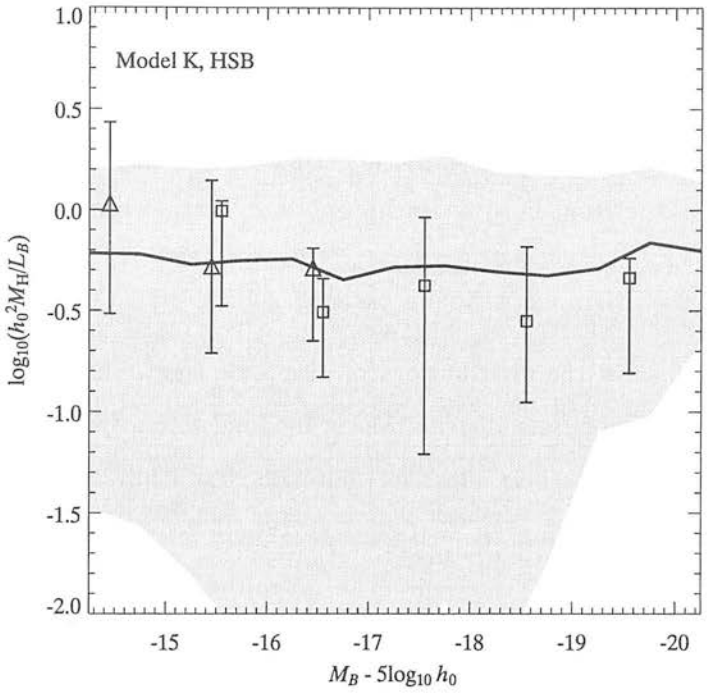

Figure 3.11: As Fig. 3.9(b) for high surface-brightness ($\mu_0 \leqslant 23$ $B-$mag arcsec^{-2}) disc galaxies only.

fractions of HSB galaxies predicted by the model agree well with the observations, but the correlation is too steep and has a much smaller scatter than observed. However, the comparison should be treated with caution because the contribution from molecular gas is not well quantified and is known to vary considerably with morphology, surface brightness, etc.

The effect of the surface brightness bias is illustrated in Fig. 3.11. This plot is equivalent to Fig. 3.9(b), but I now restrict the sample to those galaxies with $\mu_0 \leqslant 23$ $B-$mag arcsec^{-2}. At a given magnitude, this HSB sample has a median gas fraction significantly less than the total sample and reproduces the observed values well. Although the correlation with magnitude is rather flat, the overall fit is considerably improved. This is one example of how surface brightness selection effects can bias our view of the universe, so that care must be taken before rejecting models which appear, at first sight, to be in disagreement with observations.

3.3.3 Bivariate brightness distribution

de Jong & Lacey (2000) have measured the distribution of disc scale lengths as a function of luminosity for a complete sample of ~1000 Sb-Scd galaxies with well-defined selection limits (Mathewson, Ford & Buchhorn 1992, Mathewson & Ford 1996). The data are well-described by a Schechter function in luminosity, combined with a lognormal distribution of scale-sizes.

Figs. 3.12–3.15 show the distributions of disc scale lengths in bins of absolute I-band magnitude. The bivariate distribution of de Jong & Lacey is plotted as a solid histogram with error bars, their effective (half-light) radii having been converted to exponential scale lengths using $R_e = 1.678R_d$. Where no galaxies were detected in a given bin, their upper limit is given. The magnitudes are for the disc only, and are corrected to face-on, dust extincted values. The model curves include only galaxies with $(B/T)_B \leqslant 0.3$; this is slightly more restrictive than the definition of disc-dominated galaxies given earlier, to account for the fact that the de Jong & Lacey sample includes only galaxies of Hubble type Sb and later. I also plot the predictions of the semi-analytic model of CLBF (dotted histogram; results taken from de Jong & Lacey). The model curves are quite well-fit by lognormal distributions; as expected from simple arguments:

1. for an exponential disc, the scale length is given by $R_d = j_{disc}/2V_c$;
2. if the angular momentum of the baryons is conserved then $j_{disc} = j_{halo}$;
3. for a halo with an isothermal sphere profile $E = -M_{halo}V_c^2/2$, so the spin parameter (equation 2.14) reduces to $\lambda = j_{halo}V_c/GM_{halo}$;
4. the spherical collapse model gives $V_c^3 \simeq 10GM_{halo}H(z)$ for a halo of mass M_{halo} formed at redshift.

Combining these and using the Tully–Fisher relation, $L \propto V_c^\alpha$, the scale length of a galaxy disc can be written as $R_d \propto \lambda L^\beta$, where $\beta = -1/\alpha$, for a given formation redshift, z. Since the spin parameters of haloes are expected to have a log-normal distribution, independent of the other properties of the galaxy, the scale length distribution

is also expected to be log-normal, with the same width $\sigma_{\ln R} \simeq 0.5$. This is already a problem for this simple model, since the observed width is much smaller ($\sigma_{\ln R} \simeq 0.3$).

In Fig. 3.12 I compare the bivariate distributions for models K (solid line) and B (dotted line). Model B clearly predicts a much wider distribution of scale lengths at a given luminosity than is observed. In fact, the model distribution is even wider then the prediction of the CLBF model, because of the wider spin distribution measured in the N-body simulations. Independent Gaussian fits to the distributions in Fig. 3.12 yield widths in $\log_{10} R_{\rm d}$ of 0.6-0.8, in agreement with the width of the spin distribution measured in the simulations. A particular problem for this model is the excess of large scale length (LSB) galaxies in the model. de Jong & Lacey suggested that a star formation threshold would act to suppress star formation in large scale length discs, thus bringing down their numbers. The solid line in Fig. 3.12, which is for model K, shows that this is indeed the case: the number of galaxies with scale lengths in the range 2–3 kpc is reduced in all four luminosity bins, although the formal widths of the distributions remain almost unchanged, because of the large populations of small scale-length discs.

The correlation between the median scale length and luminosity can be seen more clearly in Fig. 3.16. A linear fit to the median values gives a relationship of the form $R_{\rm d(med)} \propto L^{-\beta}$, with $\beta = -0.31$, steeper than the observed slope of $\beta = -0.21$.

Models A, B and K all assume a flat rotation curve for the dark matter and neglect the self-gravity of the disc itself. Both of these assumptions affect the scale lengths of galaxies. In Fig. 3.13 I compare the scale length distribution for model K (dotted line) with those for models G (dashed line) and N (solid line), which include the corrections described in section 3.2.3, in the first case for disc self-gravity only, and in the second for the combined effects of self-gravity and an NFW dark matter profile. Current thinking suggests that this latter model should be the most realistic. However, it is clear from Fig. 3.13 that both models G and N produce discs with scale lengths that are too small on average, the difference between the NFW profile and the isothermal sphere being of secondary importance to the self-gravity of the disc. However, the slope of the size–luminosity correlation for model N, $\beta = -0.19$, is actually much closer to the observed value. The difference can be clearly seen in Figs. 3.16 (model K) and 3.17 (model N).

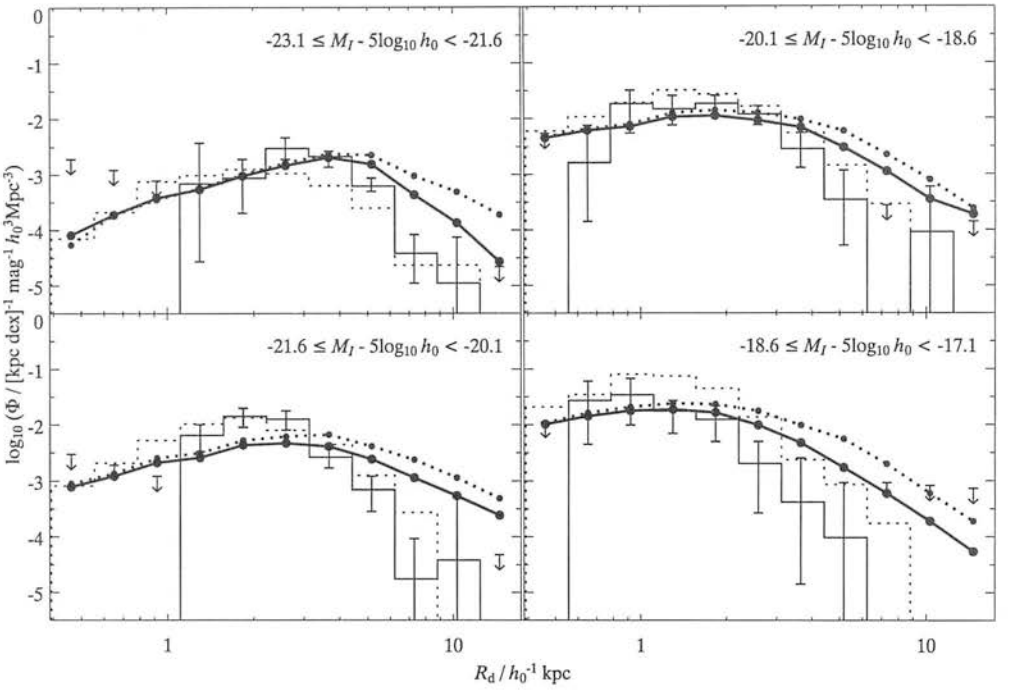

Figure 3.12: The bivariate distribution over scale length and I-band magnitude for models B (dotted line) and K (solid line), showing the effect of adding the star formation threshold. The solid histogram with error bars and the upper limits are the distribution derived by de Jong & Lacey (2000) and the dotted histogram is the prediction of a semi-analytic model (see text for details).

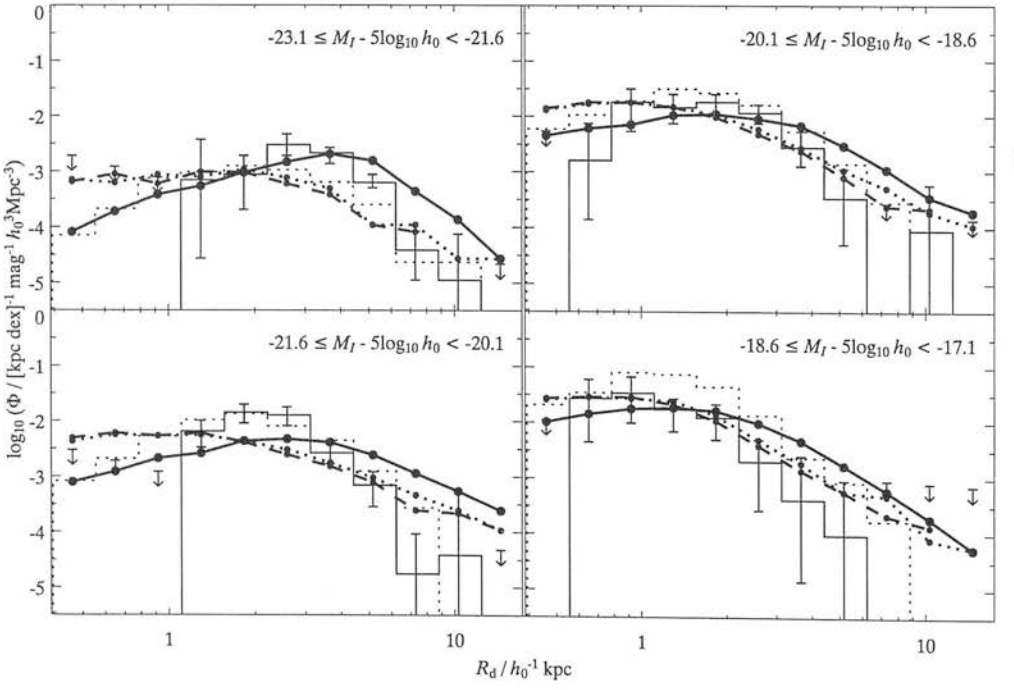

Figure 3.13: As Fig. 3.12 for models K (solid line), G (dashed line) and N (dotted line), showing the effect on the disc sizes of the self-gravity of the discs (model G) and, a dark matter halo described by the NFW profile, rather than a singular isothermal sphere (model N).

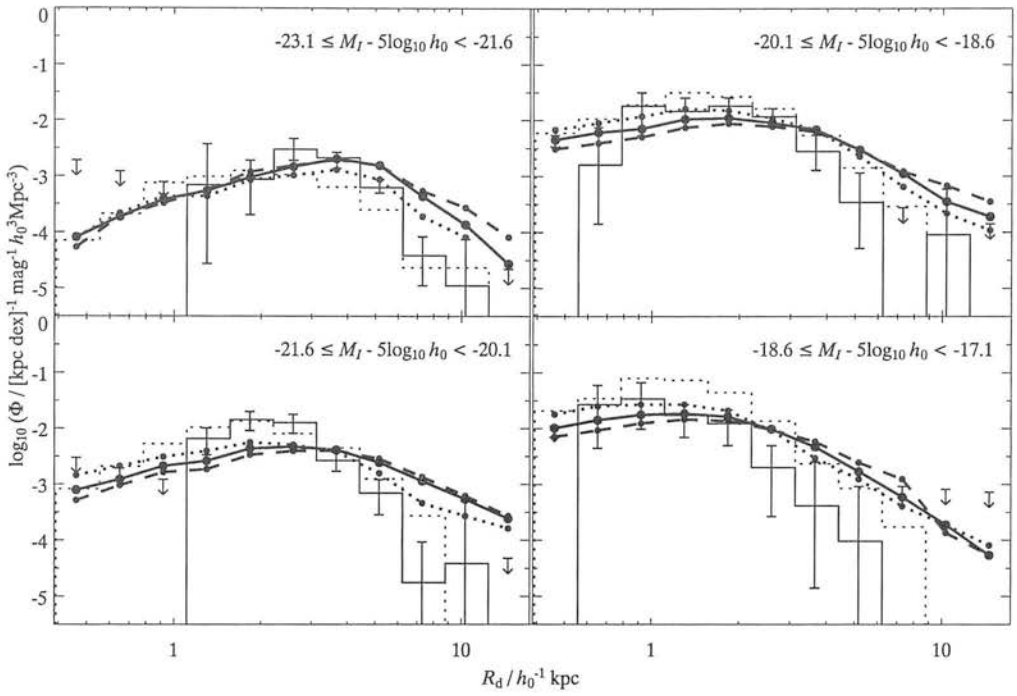

Figure 3.14: The effect on the bivariate brightness distribution of varying the feedback efficiency, $V_{\rm hot}$. The heavy solid line is for model K, while the remaining lines are for stronger ($V_{\rm hot} = 300$ km s^{-1}; heavy, dashed line) and weaker ($V_{\rm hot} = 100$ km s^{-1}; heavy, dotted line) stellar feedback. The solid histogram is the observational measurement of de Jong & Lacey (2000) and the dotted histogram is the prediction of a semi-analytic model (see text for details).

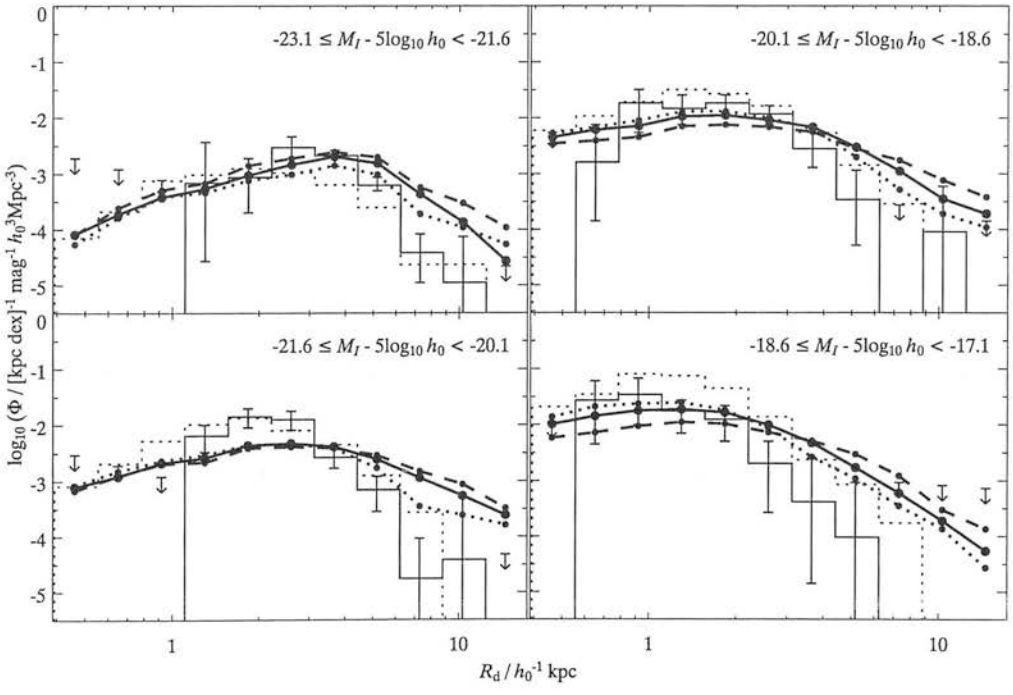

Figure 3.15: As Fig. 3.14, showing the effect of varying the feedback efficiency, $\alpha_{\rm hot}$. Again, the heavy solid line is for model K. The remaining lines are for steeper ($\alpha_{\rm hot} = 4$; heavy, dashed line) and shallower ($\alpha_{\rm hot} = 1$; heavy, dotted line) dependence of feedback efficiency on the halo circular velocity.

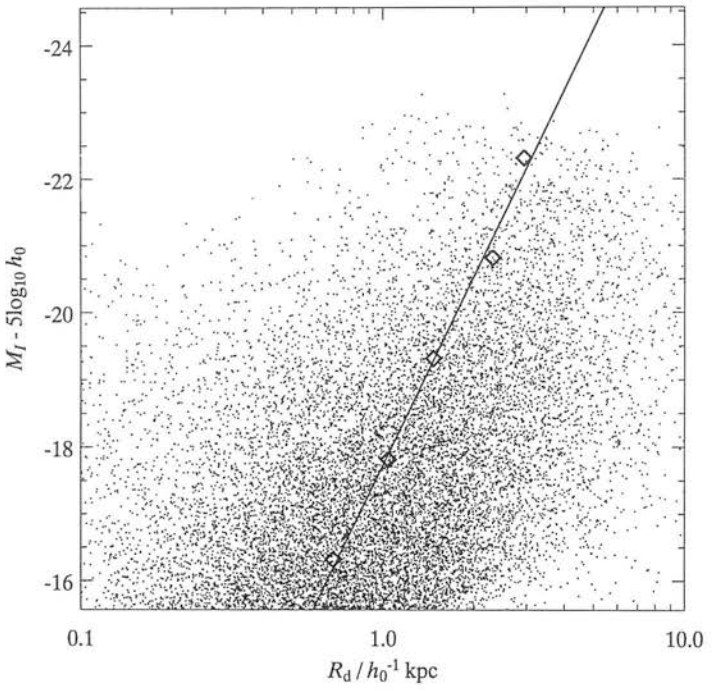

Figure 3.16: Disc scale length vs. I-band magnitude for disc-dominated galaxies in model K. The points are individual galaxies, while the diamonds mark the median scale length at each magnitude. The solid line is a linear fit to the median scale length as a function of magnitude (excluding the brightest point, since there are only a few galaxies in this bin and they are unusually compact).

Of the parameters of the basic model, the one that has the largest effect on the bivariate distribution (that is, other than simply changing the number of galaxies at a given luminosity) is stellar feedback. In Figs. 3.14 and 3.15 I show the effect on the distribution for model K of varying $V_{\rm hot}$ and $\alpha_{\rm hot}$, respectively. The main effect of increasing the strength of feedback in small haloes, either through increasing $V_{\rm hot}$ or $\alpha_{\rm hot}$, is to delay star formation, so that there is more gas available to form stars in large scale-length discs at later times. The effect is, however, small.

3.3.4 Revisiting the Tully–Fisher relation

In chapter 2 I used the Tully–Fisher relation (TFR) as a general constraint on the galaxy formation model. The only selection criterion I applied was that the galaxies used to define the relation had to be actively forming stars. Observationally, however,

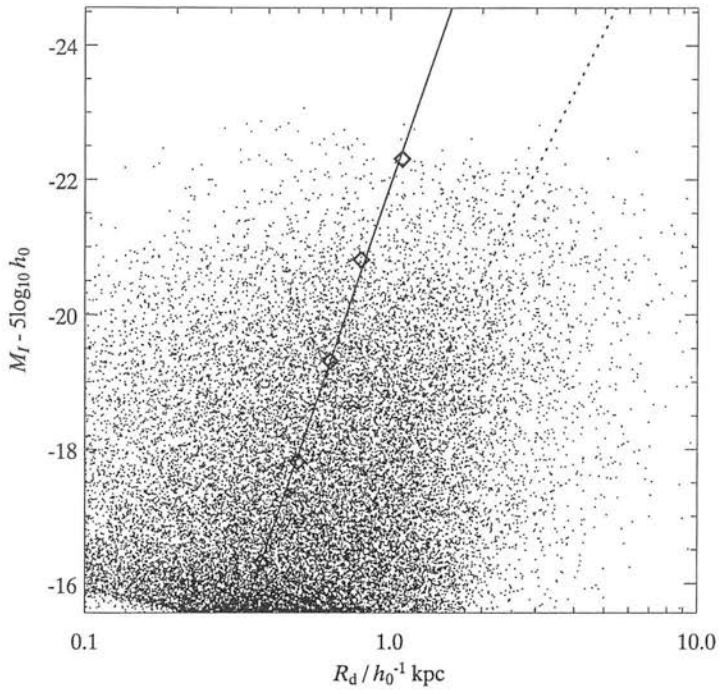

Figure 3.17: As Fig. 3.16 for disc-dominated galaxies in model N.Again, the brightest magnitude bin is excluded from the fit. For comparison, the dotted line shows the fit from Fig. 3.16.

the TFR is a relationship between the luminosities of *spiral* galaxies and the rotational velocities of their discs, which are not necessarily the same as the virial velocities of their parent haloes. In Fig. 3.18 I compare the TFRs for models K and N. The velocities plotted are the maximum rotational velocities of the discs, which are the same as the virial velocities of the haloes in model K, but in model N are calculated assuming an NFW profile for the dark matter and including the self-gravity of the discs. This leads to discs that rotate at roughly twice the expected speed, which is consistent with the results of the previous section, where I showed that the same model predicts discs that are too small at a given luminosity. This seems to be a serious problem for models that assume a centrally-concentrated halo profile. Given that self-gravity is undoubtably important in most galaxies, the only way to match the observed sizes and rotational velocities of galaxies would seem to be to abandon the centrally-concentrated dark matter profile in favour of one that is flatter in the central regions.

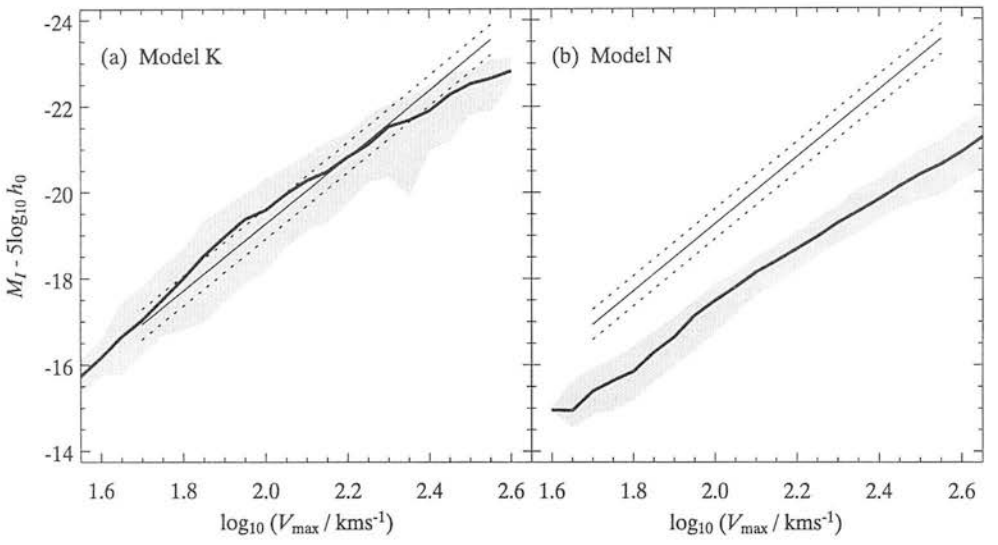

Figure 3.18: The I-band Tully–Fisher relation for (a) model K and (b) model N, showing the combined effect of including the self-gravity of the disc and an NFW profile.

3.3.5 Sizes of stellar discs

One of the more interesting predictions of the model is that spiral galaxies should have a well-defined outer edge to their star forming discs, corresponding to the outer critical radius, $R^{(2)}_{\rm crit}$. This behaviour is indeed observed in many edge-on spirals (e.g. Kregel, van der Kruit & de Grijs 2002), where a truncation of the exponential light profile is seen at radii typically several times the exponential scale length. In Fig. 3.19 I compare values of $R^{(2)}_{\rm crit}$ for galaxies in model K with the observed cut-off radii of Kregel et al. The observational data is for a small sample of galaxies and has large error bars, but it seems to show a slight tendency for $R^{(2)}_{\rm crit}/R_{\rm d}$ to increase toward smaller values of $R_{\rm d}$. This trend is mirrored in the model points, but the slope is much steeper and there are no large scale-length galaxies with large relative truncation radii.

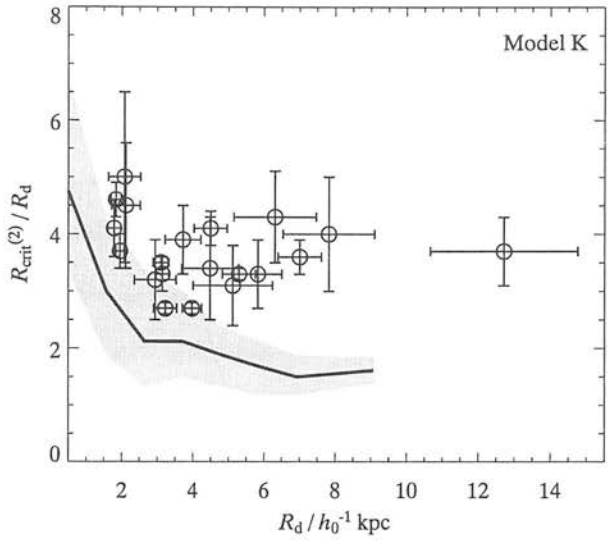

Figure 3.19: Outer critical radii of galaxies in the simulations compared to the observed distribution of truncation radii in spiral galaxies. The solid curve marks the median value of $R_{\rm crit}^{(2)}/R_{\rm d}$ and the shaded area denotes the 10 and 90% quartiles of the distribution. Circles with error bars are for the sample of Kregel, van der Kruit & de Grijs (2002).

3.4 Consequences of thresholds for galaxy formation

3.4.1 Star formation thresholds: an alternative to feedback?

In Fig. 3.20 I compare the effect of adding a star formation threshold with the separate effect of increasing $\alpha_{\rm hot}$ (recall that this gives the slope of the dependence of feedback efficiency on circular velocity). Starting from the basic model, with $\alpha_{\rm hot} = 1$ (i.e. a weak dependence on circular velocity), the effect of increasing $\alpha_{\rm hot}$ to its fiducial value of 2 is to flatten the faint-end slope. Adding a star formation threshold (model K) has much the same effect. Note that model K also includes morphological evolution, which is not present in the basic model, and has a different merger timescale, but neither of these has a large effect on the luminosity function (see Fig. 2.19a for the effect on the luminosity functions of changing $f_{\rm mrg}$). The combined effect of a star formation threshold and the fiducial feedback strength brings the faint end slope down even more, but results in too few galaxies at intermediate luminosities. This is illustrated in Fig. 3.21. The same is true of models in which it is the parameter $V_{\rm hot}$ that is altered. So it seems that weak

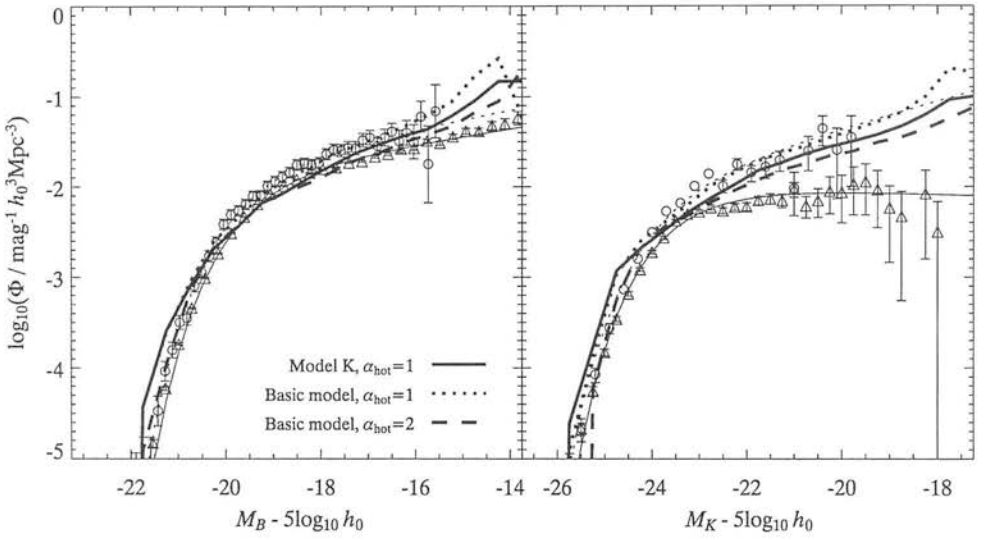

Figure 3.20: A comparison of the effects on the B- and K-band luminosity functions (LFs) of feedback and star formation thresholds. The solid curve in each panel is for a variant of model K that has weak feedback in low-mass haloes ($\alpha_{\rm hot} = 1$). The remaining curves are for the basic model; the dotted curve is for a model that also has $\alpha_{\rm hot} = 1$ while the dashed curves are for $\alpha_{\rm hot} = 2$. The model LFs are plotted over observational data as in Fig. 2.18.

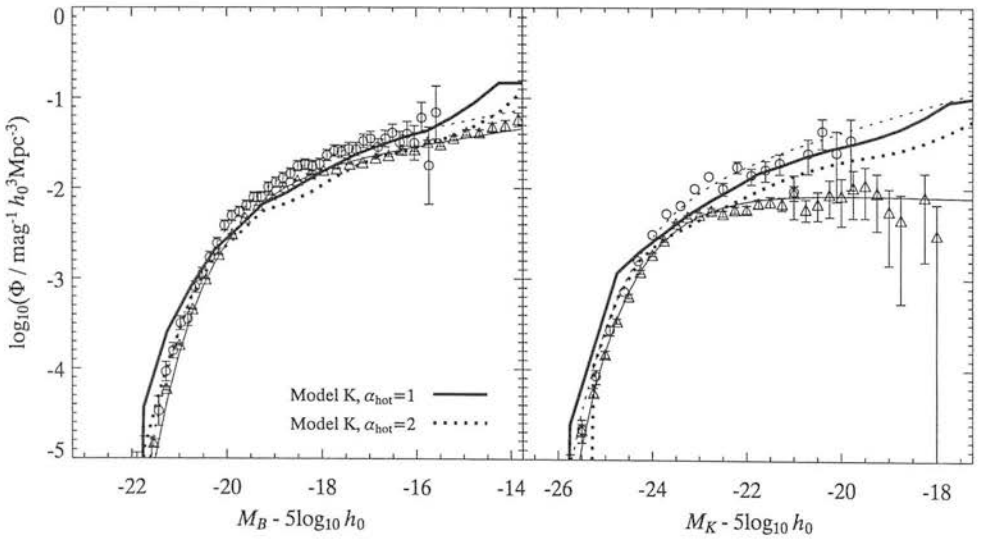

Figure 3.21: The combined effect on the B- and K-band luminosity functions (LFs) of feedback and star formation thresholds. Both of the model curves are for model K; the solid curves are for $\alpha_{\rm hot} = 1$, as in Fig. 3.20, while the dotted curves are for $\alpha_{\rm hot} = 2$, the fiducial value in chapter 2. The model LFs are plotted over observational data as in Fig. 2.18.

feedback is not only possible, but necessary, in models with thresholds. In any case, it is clear that star formation thresholds are at least as good a way of flattening the faint-end slope of the luminosity function as feedback. Unlike the feedback prescription, which is a simple scaling law with free parameters, star formation thresholds have the advantage of being observationally-motivated and the critical density, parameterised in $\alpha_{\rm K}$, is fixed by these observations. Evidence for strong feedback from star formation in 'normal' galaxies has, on the other hand, proven to be elusive.

3.4.2 Where are the baryons?

Kennicutt (1989) observed that the azimuthally-averaged gas surface density profile in spiral galaxies closely follows (to within a factor of ~ 2) the critical density for star formation predicted by gravitational stability arguments (see also Martin & Kennicutt 2001). This strongly suggests that the gas contents of spirals are primarily determined by the star formation threshold and that the star formation history must therefore be controlled by the rate of infall of fresh gas onto the disc. Thus, one might expect the star formation histories of high and low surface-brightness galaxies to respond differently to the presence of thresholds. In Figs. 3.22 and 3.23 I plot the median fraction of the baryons in each phase (hot gas, cold gas, stars) as a function of time for $L^\star$ and $L^\star/10$ galaxies, respectively. At each luminosity, I divide the galaxies into HSB ($\mu_0 \leqslant 23$ $B-$mag arcsec^{-2}, top rows) and LSB ($\mu_0 > 23$ $B-$mag arcsec^{-2}, bottom rows) samples. The left column shows the predictions of model B and the right column the predictions of model K. The upper part shows the fraction of the final baryonic mass in place, also as a function of time. Note that the corresponding samples for the two different models do not necessarily contain the same galaxies: the samples are defined on the basis of the present-day properties predicted by each model separately.

The difference between the two models is most pronounced in low-luminosity galaxies. In model B, the cold gas fraction in these galaxies is kept low by the action of feedback, which ensures that the dominant fraction of the baryons is kept in the hot phase. This is particularly true of HSB galaxies, in which the star formation timescale is small. In LSB galaxies the timescale for depletion of the cold gas reservoir is longer, but feedback is more efficient so the fraction turned into stars is smaller and a larger

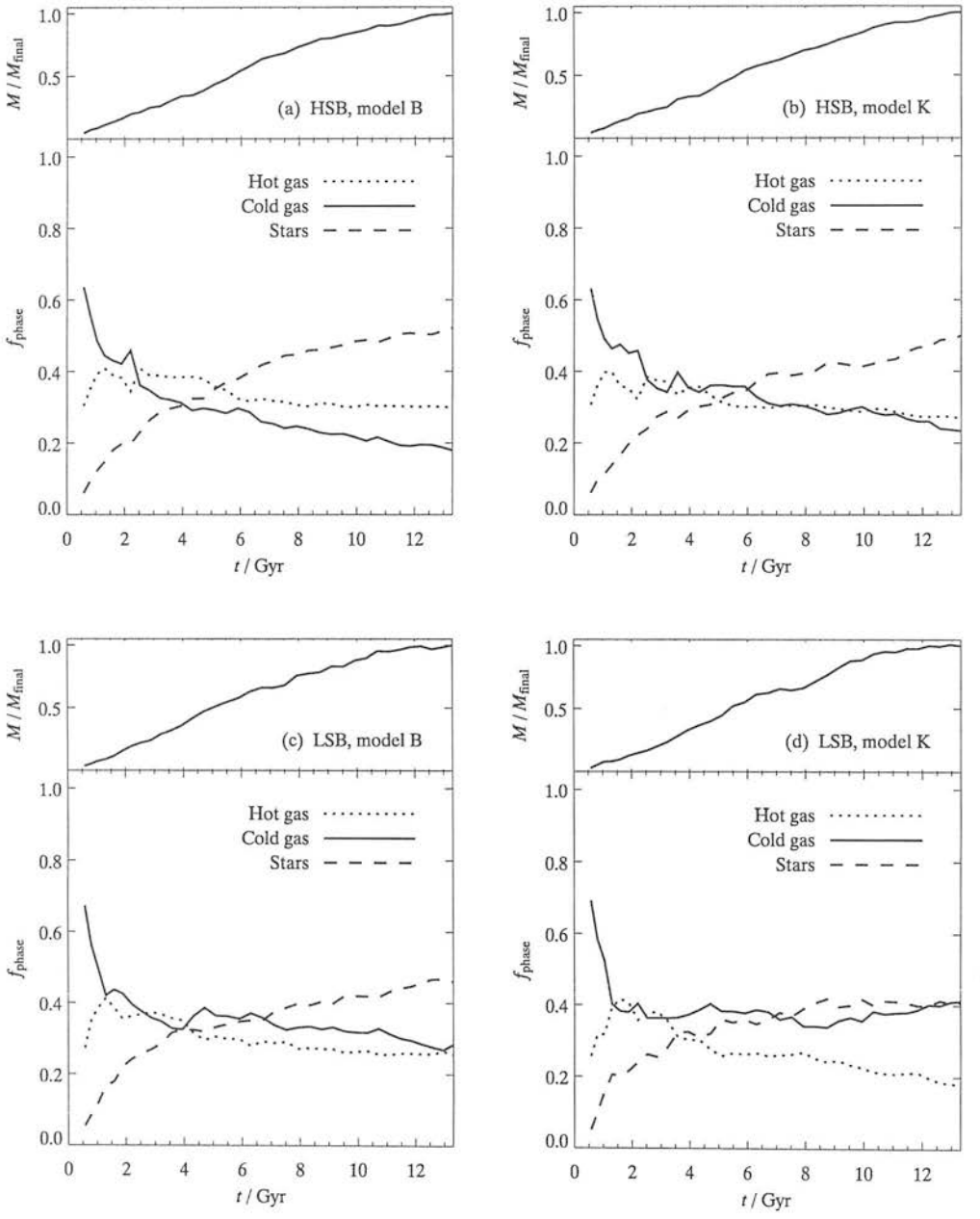

Figure 3.22: The evolution of the hot and cold gas reservoirs and the mass of long-lived stars for the most massive progenitors of present-day L^* galaxies. Top row: high surface-brightness ($\mu_0 \leqslant 23$ $B-$mag arcsec^{-2}) galaxies in model B (left) and model K (right). Bottom row: low surface-brightness ($\mu_0 > 23$ $B-$mag arcsec^{-2}) galaxies in the same two models. The main part of each panel shows the median fraction in each phase for galaxies in the range $M^* \pm 1$ mag, while the top parts show the fraction of the final baryonic mass in the most massive progenitor as a function of time.

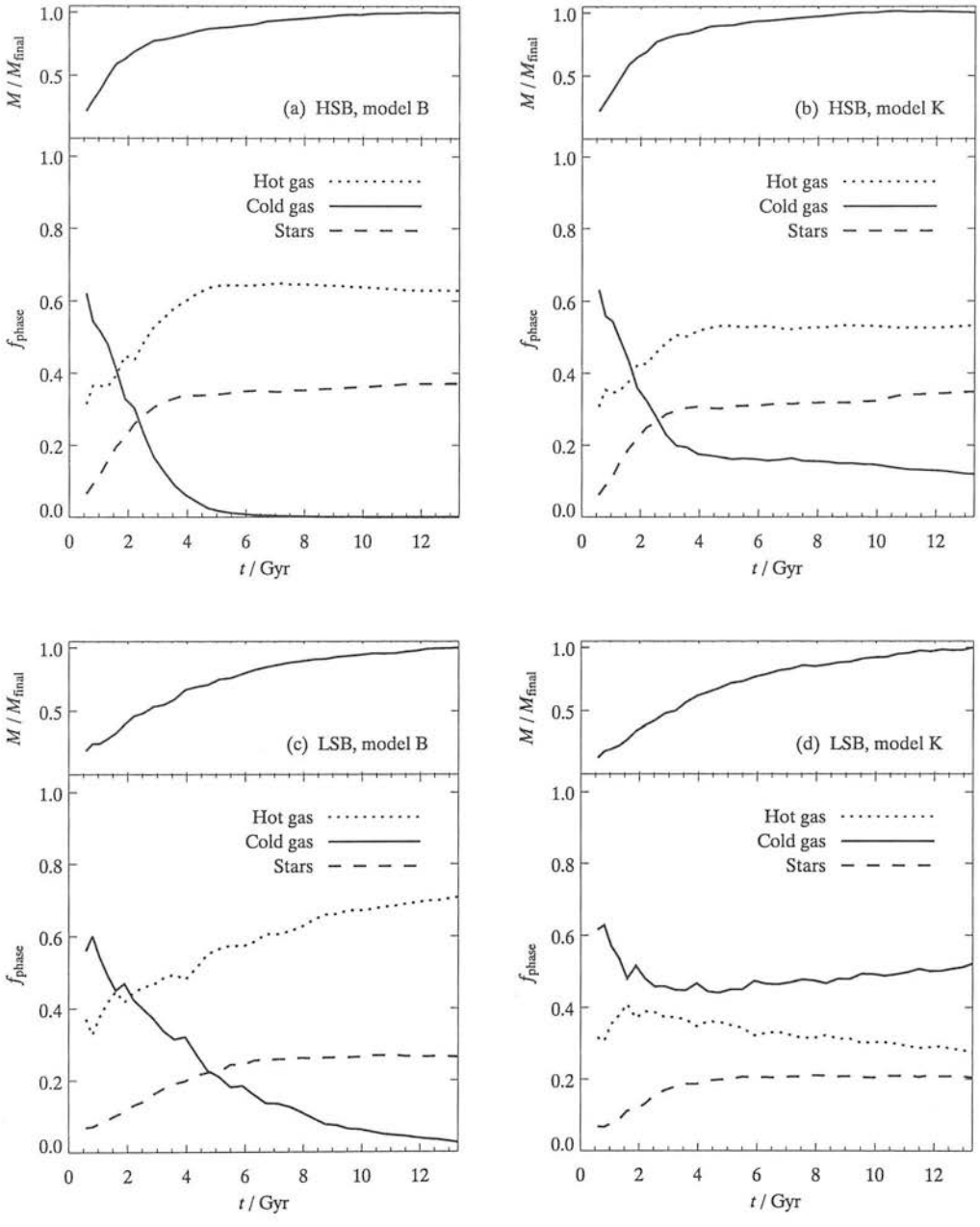

Figure 3.23: As Fig. 3.22 for $L^{\star}/10$ galaxies.

fraction is reheated to the hot phase. In both cases, feedback limits the fraction of baryons that are turned into stars but the cold gas fractions are inconsistent with observations, leading to the trend seen in section 3.3.2. In model K, the cold gas fraction in low-luminosity galaxies is maintained at a more or less constant level for most of the history of the galaxies, through the action of the star formation threshold. The greatest difference is seen in LSB galaxies: model K predicts that, contrary to the predictions of model B, the greatest fraction of the baryons in low-luminosity LSB galaxies should be in the form of a cold gas disc. The effects of thresholds are not restricted to low-luminosity galaxies: even in L^* LSB galaxies, model K predicts that cold gas should make up a similar fraction of the total baryonic content to stars.

3.5 Discussion

I have shown how the observed gas contents of spiral galaxies can be conveniently explained by the existence of a threshold density, below which widespread, massive star formation is not possible. Such thresholds are expected from gravitational stability arguments and have been observed directly in external galaxies (Kennicutt 1989; Martin & Kennicutt 2001). The possibility of a non-linear star formation law is often neglected in galaxy formation models, but star formation laws in which the timescale for star formation in galaxies is simply proportional to the dynamical times of their discs, as observed at high densities (Kennicutt 1998), predict gas contents for low-luminosity galaxies that are too low to be consistent with observations. An empirical scaling of the star formation timescale with circular velocity, as adopted by some authors, predicts the correct correlation between gas content and luminosity, but has no obvious physical motivation, unlike the non-linear law adopted here, which has both good theoretical motivation and observational evidence in its favour.

de Jong & Lacey (2000) suggested that star formation thresholds might play an important role in shaping the bivariate brightness distribution, a 2-dimensional generalisation of the LF that includes the sizes of galaxies. LFs are one of the most important constraints on galaxy formation models, but they provide only a partial description of the galaxy population, for reasons discussed in chapter 4. A full description of the

galaxy population requires a distribution over at least two of three parameters: luminosity, surface brightness and scale size; and a sample that is complete to well-defined selection limits in the chosen parameters. A bivariate distribution over any two of these parameters is not biased by surface-brightness selection effects and also provides a stronger constraint on galaxy formation models.

Models with no star formation threshold predict a wider distribution of disc scale lengths at a given luminosity than is observed. Star formation thresholds act to suppress star formation in low surface-density discs, leading to fewer large scale-length galaxies at a given luminosity. The effect is, however, not large enough to bring the predicted distribution into agreement with the observations and the width of the distribution is still to large. To correct this, a mechanism is also required to reduce the numbers of small scale-length discs.

The scale lengths of discs are strongly affected by their own self-gravity. Models in which this effect is taken into account predict the correct slope of the correlation between median scale length and luminosity, which is not the case for models that assume massless discs. The latter give a slope consistent with theoretical expectations from the simple model outlined in section 3.3.3, suggesting that self-gravity may be the cause of the difference between the expected slope and that observed. However, they also predict galaxies that are, on average, too small for their luminosity. Another consequence of this that the model TFR is offset from the observed relation by a factor of order 2 in circular velocity. It is difficult to see how the predictions of the model, particularly the zero-point of the TFR, can be reconciled with observations unless the circular velocities of discs are close to the virial velocities of their haloes. As pointed out by Mo & Mao (2000), this would require haloes to be less centrally-concentrated than those seen in N-body simulations. The existence of a central cusp in the density profiles of galaxy haloes is a matter of controversy; in particular, the rotation curves of LSB dwarf galaxies have been found by many authors to be better fit by profiles with constant density cores (de Blok & McGaugh 1997; Swaters, Madore & Trewhella 2000; de Blok & Bosma 2002) than by NFW-type profiles. Such profiles have been given little attention in the field of galaxy formation modelling, so it would be interesting to know if they are capable of producing discs of the correct size while preserving the slope of

the correlation with luminosity predicted for self-gravitating discs.

3.6 Conclusions

I have investigated the effects of a non-linear star formation law on the properties of disc galaxies, by incorporating the star formation threshold of Kennicutt (1989) into the phenomenological galaxy formation model set out in chapter 2. The existence of star formation thresholds affects the evolution of galaxies in two ways: by limiting the amount of gas that can be turned into stars and by defining the star-forming region of the disc, thus determining the amount of gas considered 'available' for star formation. Some galaxies are unable to form any stars because their gas discs are entirely sub-critical. I have shown how star formation thresholds can explain the observed gas contents of spiral galaxies, provided selection effects in surface-brightness are accounted for. Models with thresholds also reproduce the observed correlations between gas content and both luminosity and surface brightness.

Star formation thresholds go some way towards reducing the width of the scale length distribution, by limiting, and in some cases preventing, widespread star formation in low surface-density discs. This confirms the speculative suggestion of de Jong & Lacey (2000). The problem of reducing the numbers of small scale-length discs, however, remains. The observed scale lengths and circular velocities of galaxy discs cannot be reproduced by models in which the dark matter profile has a central cusp; the combined effects of the cuspy density profile and the self-gravity of the disc produce discs that are too small for their luminosities and rotate too rapidly to lie on the observed TFR. The closest fit to the bivariate brightness distribution of galaxy discs is achieved by making the basic assumptions of a massless disc and an isothermal sphere halo profile.

Chapter 4

Surface brightness selection effects

I this chapter I apply the model developed over the previous two chapters to the question of the role played by surface brightness selection effects in shaping the observed galaxy luminosity function. I do this by constructing mock surveys using the simulation boxes and selecting galaxies on the basis of their *isophotal* magnitude — the integrated flux above some limiting isophote. Samples selected in this way miss many faint, low surface-brightness galaxies, even though their total fluxes put them within the formal magnitude limit. The conclusion I reach is that it is possible to reconcile even models with intrinsically very steep faint-end slopes with the observed luminosity function, which removes one of the motivations for strong stellar feedback in galaxy formation models.

4.1 Introduction

The galaxy luminosity function (LF) is one of the most important constraints on galaxy formation models. Models in which feedback from star formation is less efficient at reheating gas in low-mass haloes generally predict a steeper faint-end slope ($\alpha \simeq -2$) to the LF (Kauffmann, White & Guiderdoni 1993; Cole et al. 1994; CLBF), in apparent conflict with the observed slope ($\alpha \simeq -1.20$ to -1.25; see table 1.1). However, the LF gives an incomplete description of the galaxy population, for the following reason: all

observational samples have, in addition to the formal magnitude (or in some cases diameter) limit, an isophotal (surface brightness) limit, which biases them against the inclusion of low surface-brightness galaxies of all luminosities. An extreme example of this is Malin 1 (Bothun et al. 1987) which has a total apparent magnitude of $B = 14.2$ and is, in fact, the most luminous galaxy in the local universe, but is excluded from many samples claiming completeness to fainter magnitude limits because its low surface brightness means that most of the flux is below the limiting isophotes of such samples. A LF derived from a magnitude-limited sample is thus only valid over some range in surface brightness at a given luminosity. It has been suggested (Cross et al. 2001) that the different isophotal limits of surveys may be responsible, at least in part, for the different values of the Schechter parameters measured from different surveys. In fact, the disagreement between the observed faint-end slope and the predictions of galaxy formation models is considerably less serious now that it was a decade ago, when the best data suggested a slope $\alpha \lesssim 1$; Loveday et al. 1992).

Galaxy samples are usually taken to be complete down to some limit in apparent magnitude (or, in some cases, angular diameter). The apparent magnitude of a galaxy is related to its absolute magnitude by $m = M + 5\log_{10} d_{\rm L} + 25$, where $d_{\rm L}$ is the luminosity distance to the galaxy, measured in Mpc. So, naturally, observational samples are dominated by intrinsically bright galaxies, since these can be seen at much greater distance than galaxies with fainter absolute magnitudes. This bias must be accounted for when deriving quantities such as the space density of galaxies as a function of absolute magnitude, i.e. the LF. The simplest estimator for the LF is the *maximum-volume* ($V_{\rm max}$) estimator. For each galaxy, i, in a sample, an equivalent density is calculated based on the distance, $d_{\rm max}$, at which the galaxy in question would be expected to drop out of the sample, on the basis of the formal selection criteria. The galaxies are then binned in absolute magnitude to give an estimate of the LF:

$$\Phi(M)\,\mathrm{d}M = \sum_i \frac{1}{V_{\rm max}^i}, \tag{4.1}$$

where the sum is over all galaxies for which $M - \mathrm{d}M/2 \leqslant M_i < M + \mathrm{d}M/2$. For a magnitude-limited sample, the volume over which a galaxy of luminosity L is seen is

given by

$$\begin{aligned} V_{\rm max}(L) &\propto d_{\rm max}^3 \\ &\propto L^{3/2}, \end{aligned} \tag{4.2}$$

where $d_{\rm max}$ is the distance at which the galaxy drops out of the sample. The $V_{\rm max}$ estimator relies on the assumption of a uniform galaxy distribution and so is rarely used in practice because the true galaxy distribution is strongly clustered. More advanced estimators give results that are independent of inhomogeneities in the distribution of galaxies: the two most commonly used are the STY (Sandage, Tammann & Yahil 1979) and *stepwise maximum likelihood* (Efstathiou, Ellis & Peterson 1988) estimators. Both are maximum-likelihood techniques but, while the former requires a functional form (such as the Schechter function) to be assumed, the latter is independent of the shape of the LF.

Equation 4.2 is only strictly applicable to a point source, for which the *total* flux can be measured. In practice, for resolved sources such as galaxies, a true flux-limited sample does not — and indeed cannot — exist, because all surveys have a limiting isophotal level and any flux distributed below this level is missed. The consequence of this is that flux-limited (and, to a lesser extent, diameter-limited) samples are biased against the inclusion of LSB galaxies, *regardless of their total flux.* To gain a complete description of the galaxy population from a magnitude-limited sample, it is necessary for the sample to be complete to well-defined magnitude ($m_{\rm lim}$) and isophotal ($\mu_{\rm lim}$) limits and to have 2-dimensional photometry for all galaxies, from which a distribution over two parameters can be derived. McGaugh, Bothun & Schombert (1995) give the full expression for the visibility of a galaxy with central surface brightness μ_0 and scale size h:

$$\begin{aligned} V_{\rm max}(h,\mu_0) &\propto L_{\rm iso}^{3/2} \\ &\propto h^3 10^{-0.6(\mu_0-\mu_{\rm lim})}[f(R_{\rm iso})]^{3/2}, \end{aligned} \tag{4.3}$$

where $L_{\rm iso} = f(R_{\rm iso})$ is the flux measured above the limiting isophote and $R_{\rm iso}$ is the isophotal radius (note that I measure the surface brightness relative to the isophotal limit, $\mu_{\rm lim}$, rather than the Freeman value, $\mu_0^\star = 21.65$ $B-$mag arcsec^{-2}, as they did).

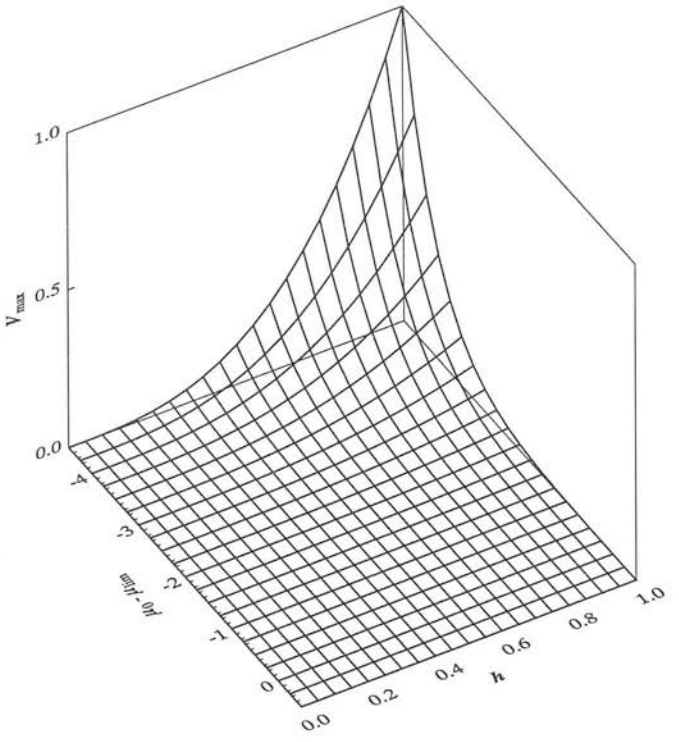

Figure 4.1: The visibility of galaxies in a flux-limited sample, as a function of central surface brightness, μ_0, and linear size, h. Plotted is the effective volume, $V_{\rm max}$, over which a galaxy with the given properties will be seen in a sample limited by the apparent isophotal magnitude. The units of $V_{\rm max}$ are arbitrary, as are those of h, and μ_0 is measured relative to the isophotal limit, $\mu_{\rm lim}$. $M^\star$ is the usual Schechter function parameter.

This is plotted in Fig. 4.1, from which it is clear that samples with isophotal magnitude limits are biased against the inclusion of galaxies with low surface brightnesses, as well as those with small angular extents. Diameter-limited samples are known to be less strongly biased against the inclusion of LSB galaxies, but most large galaxy surveys are magnitude-limited.

One way to ensure a fair comparison between observations and the predictions of galaxy formation models is to compare to a *bivariate* brightness function, such as that of de Jong & Lacey (2000), with which I compared the predictions of the model in chapter 3. As well as providing a stronger constraint on the model, a bivariate brightness distribution is not affected by surface brightness selection effects. Unfortunately, large samples of galaxies complete to well-defined selection limits are rare. An alternative

approach is to take the predictions of a model and apply the same selection criteria as to the observed sample. The two approaches are complimentary; while the latter doesn't give as strong a constraint, it at least prevents models being ruled-out prematurely because the observational constraints are taken at face value. An example of this is the need for strong stellar feedback in galaxy formation models to explain the faint-end slope of the LF. In this chapter, I use mock redshift surveys to show how surface-brightness selection effects can explain the faint-end slope of the LF without the need for strong feedback.

4.2 Modelling selection effects

4.2.1 Constructing a survey volume

Mock survey volumes are constructed by lining up simulation boxes along the line-of-sight between an observer at $z = 0$ and some high redshift. The first useful output from the simulations is at $z \simeq 8.5$ so, in practice, this fixes the maximum redshift of the mock surveys. At a given redshift, z, the redshift element corresponding to a comoving length ΔL is

$$\Delta z = \frac{H(z)}{c} \Delta L \tag{4.4}$$

(see appendix C), where $H(z)$ is given by equation (1.13). The simulation boxes have a fixed comoving length of $32h_0^{-1}$ Mpc. The redshift of each box is found by summing these contributions along the line of sight from the observer, starting from the centre of the first box ($z = z_0$) and working out in steps of $\Delta L/2$. This requires ~ 200 boxes along the line of sight, many more than the actual number of snapshots ($\simeq 40$) available of each simulation box, so it is neccessary to interpolate the galaxy properties between the snapshots. The transverse comoving length, $\Delta L_{\rm T}$, corresponding to an angle $\Delta\psi$ on the sky at redshift z, is

$$\begin{aligned} \Delta L_{\rm T} &= R_0 S_k(r)\, \Delta\psi \\ &= (1+z) D_{\rm A}\, \Delta\psi, \end{aligned} \tag{4.5}$$

where $D_{\rm A}$ is the angular diameter distance (equation C.3). Even for a survey with small angular size (a 'pencil-beam' survey), such as I consider here, this length exceeds

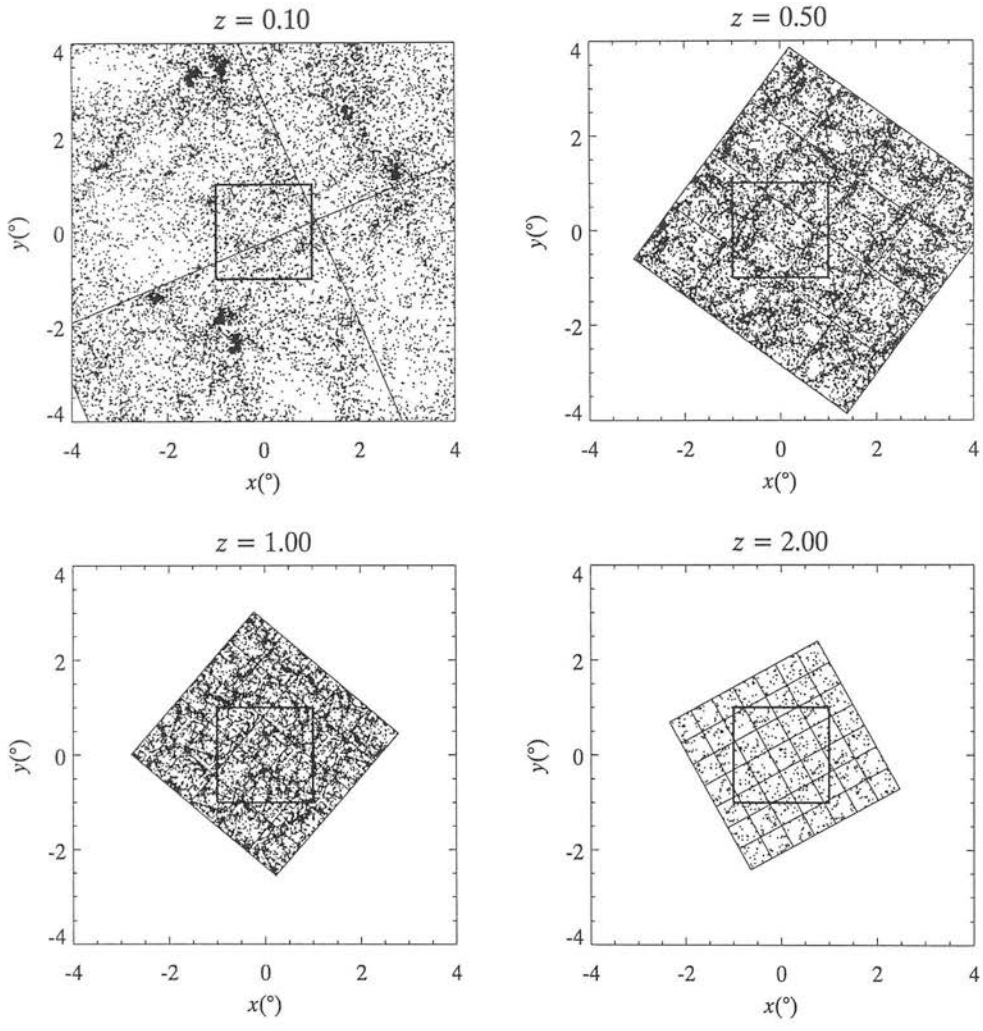

Figure 4.2: Arrangment of boxes in a mock survey volume at four different redshifts. The dimensions plotted on the x and y axes are angle on the sky, in degrees. The heavily outlined box in the centre of each panel marks the limit of a $2° \times 2°$ square survey area. Only galaxies with $B < 25$ are included.

the width of a single box at moderately high redshifts ($z \simeq 0.33$ for a $2° \times 2°$ survey, as shown in Fig. 4.2), so it is neccessary to use the same box several times at the same redshift.

Building a survey in this way clearly means that clustering along the line of sight on scales larger than the comoving box size is not taken into account. In effect, I am assuming that the boxes used are representative of the universe as a whole. In fact, the initial conditions were set up in such a way that each box has the average properties

of the universe. This is not, however, true of real survey volumes. As pointed out in chapter 2, this means that the effects of cosmic variance will be underestimated. Using the same simulation box many times, both along and perpendicular to the line of sight, clearly also introduces an artificial clustering signal. A number of steps are therefore taken to minimise contamination from such effects:

1. each box is randomly rotated through multiples of 90° about its three axes, so that different faces of the box are used each time;

2. the galaxies in each box are randomly shifted *en mass* in three dimensions, in a periodic fashion, so that galaxies that leave the box by one face enter through the opposite face;

3. the boxes are placed on a grid and the whole grid is then randomly translated by up to half the box length in a plane perpendicular to the line of sight;

4. finally, the entire grid is rotated by a random angle about an axis parallel to the line of sight.

The results of this procedure are illustrated in Fig. 4.2 for four different redshifts. Clustering of galaxies on the sky can have an effect on the LF, particularly at the faint end, because the faintest galaxies are only seen over a small volume and we want to make sure that this volume is as representative as possible of the whole simulation volume. These considerations aside, the actual distribution of galaxies on the sky is not important for this application, so I combine the five simulation boxes by constructing a separate $5° \times 5°$ survey for each simulation and simply adding the results (which is equivalent to laying them side-by-side on the sky) to create one 125 deg^2 survey.

4.2.2 Galaxy properties

As already discussed, the number of boxes required to form a contiuous volume along the line of sight is larger than the number of available snapshots. To estimate the luminosity of a galaxy at a given time, I take the simplest possible approach, performing a straightforward linear interpolation between the luminosities at the two closest snapshots. This method works for most galaxies, but a problem arises in the situation

where several galaxies exist as separate entities at one snapshot, but have merged into a single, larger galaxy at the next. There is no simple way of dealing with this problem; the simulation is set up in such a way that it is not possible to predict the properties of the progenitor galaxies immediately prior to a merger event. Instead, I interpolate between the final galaxy and its most massive progenitor. This is not very satisfactory, since it does not take account of the fact that the galaxies existed as separate entities for a fraction (in some cases most) of the time between the two snapshots. A more consistent treatment would require more information about galaxies to be stored immediately prior to each merger event.

The other piece of photometric information needed is the scale length of the disc of each galaxy. I do not interpolate this between snapshots, because the scale length is not a smoothly varying quantity but remains approximately constant throughout the lifetime of a galaxy halo (since it is determined primarily by the angular momentum of the cooling gas).

4.2.3 Number counts and redshift distribution

I take model K from the previous chapter as my fiducial model. Here I present two statistics which I use to test the method outlined above. In Fig. 4.3 I plot the galaxy number counts as a function of apparent magnitude in the B band. On the whole the predictions of the model agree well with the observed number counts in the range shown, although they are a little on the high side, especially at fainter magnitudes. The redshift distribution for galaxies brighter than $B = 19.71$ is given in Fig. 4.4. The limit was chosen to correspond to the limiting magnitude of the 2dFGRS survey, $b_{\rm J} = 19.45$, converted to the Johnson B band using equation (1.46). The distribution is somewhat broader than that of the North Galactic Pole (NGP) field of the 2dFGRS, which shows a sharp peak at the median redshift of 0.1 (I use the NGP field since the South Galactic Pole field may be affected by large-scale structure; Norberg et al. 2002). The median redshift of the model galaxies is also 0.1, but the high-z tail is more prominent in the model distribution. In both cases there are very few galaxies beyond $z \simeq 0.5$. For comparison, I also plot the redshift distribution predicted by the model $z = 0$ LF, combined with corrections for redshifting of the observing band (k-

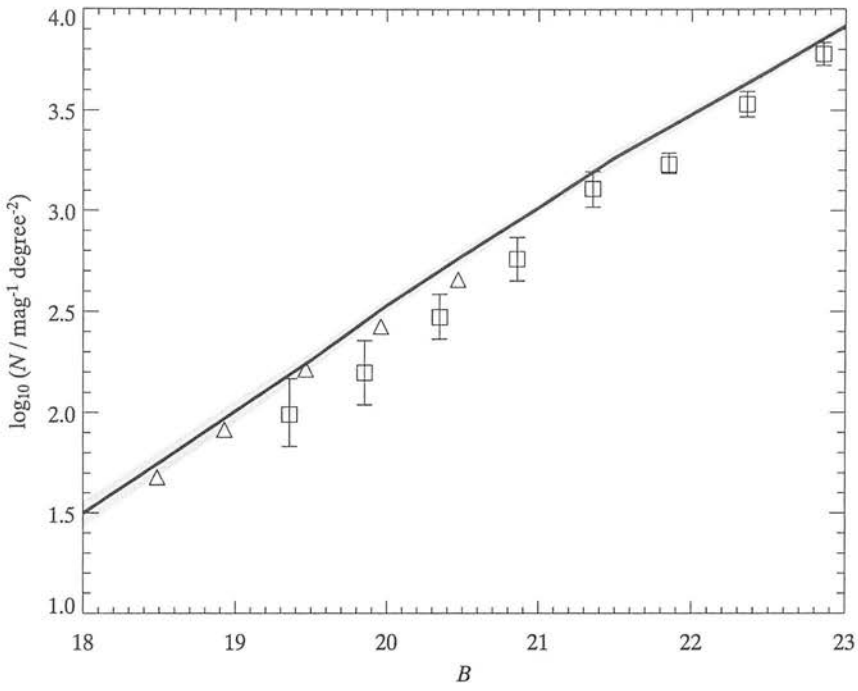

Figure 4.3: Differential number counts in the B band for the combined 125 deg^2 mock survey. The solid curve shows the predicted counts from model K and the shaded area marks the r.m.s. variance between the five individual $5° \times 5°$ fields that make up the full survey. The data are taken from two sources: triangles show the APM number counts (Maddox et al. 1990) while squares with error bars are the data of Metcalfe et al. (1991).

correction) and for evolution (e-correction). I take the mean $k+e$ correction of Norberg et al.: $k(z) + e(z) = (z + 6z^2)/(1 + 20z^3)$ (this mean correction does not account for differences with spectral type, but Norberg et al. find that this makes little difference to the reconstructed LF). The distribution of redshifts predicted using this method agrees well with that measured in the mock surveys and still overpredicts the number of galaxies at the high-redshift end of the sample, relative to the 2dFGRS.

4.2.4 Selecting galaxies

I select galaxies on the basis of their isophotal magnitudes. Recall that the model galaxies consist of two structural components:

1. A flat, circular and infinitesimally thin disc with an exponential surface brightness

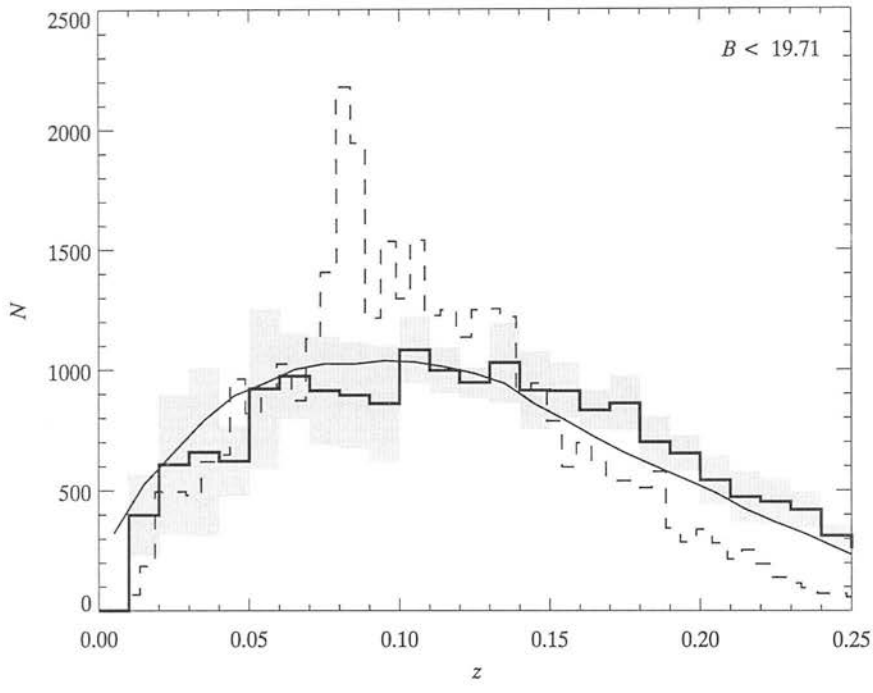

Figure 4.4: The redshift distribution for galaxies with $B < 19.71$ (corresponding to $b_J < 19.45$, the limit of the 2dFGRS) in the combined 125 deg^2 mock survey. The solid histogram gives the distribution for model galaxies while the shaded area marks the r.m.s. variance between the five individual $5° \times 5°$ fields that make up the full survey. The redshift distribution of galaxies in the North Galactic Pole field of the 2dFGRS (Norberg et al. 2002) is given by the dashed histogram, which is normalised to give the same total number of galaxies in the field. The smooth, solid curve is the redshift distribution predicted by the model LF at $z = 0$ and the mean 2dF $k + e$ correction (see text).

profile, comprising a homogeneous mixture of stars and dust.

2. A spherically symmetric bulge with an $R^{1/4}$–law light profile in projection, which is unaffected by dust extinction except in the case of recent mergers, where I again assume a homogeneous mixture of stars and dust.

The model set out in chapters 2 and 3 predicts the luminosities, but not the sizes, of spheroids. Ellipticals, however, have a fundamental plane similar to the Tully-Fisher relation for spiral galaxies. For ellipticals, this takes the form

$$\log_{10}(R_e \,/\, \mathrm{arcsec}) = \alpha \log_{10}(\sigma_0 \,/\, \mathrm{km\,s^{-1}}) + \beta(\mu_e \,/\, \mathrm{mag.\,arcsec^{-2}}) + \gamma, \qquad (4.6)$$

where r_e is the half-light, or effective, radius; μ_e is the mean surface brightness within r_e and σ_0 is the central stellar velocity dispersion. The bulges of early-type spirals are also found to lie close to the fundamental plane defined by ellipticals and lenticulars (Falcón-Barroso, Peletier & Balcells 2002). A fit to the B-band fundamental plane for a combined sample of ellipticals, lenticulars and bulges by these authors yields values of $\alpha = 1.15$, $\beta = 0.31$ and $\gamma = -8.37$. Adopting this result and using equation (D.10), the B-band fundamental plane can be written as

$$\log_{10}(R_e \,/\, \mathrm{arcsec}) = 14.09 - 2.09 \log_{10}(\sigma_0 \,/\, \mathrm{km\,s^{-1}}) - 0.56B. \qquad (4.7)$$

As an estimate of the stellar velocity dispersion I take the velocity dispersion of the parent halo. Wyse, Gilmore & Franx (1997) find that the *mean* stellar velocity dispersions of bulges are correlated with the virial velocities of their parent haloes: $\langle v^2 \rangle^{1/2}_{\mathrm{bulge}} \simeq 0.7 V_c$, although individual galaxies deviate from this relation by as much as a factor of 2 and the central velocity dispersion, σ_0, will generally be somewhat higher.

I assume an $R^{1/4}$–profile (equation D.7) for spheroids of all sizes. In reality, bulges in spiral galaxies often have profiles closer to exponential, which may suggest that they are closely related to the disc component (Wyse 2000). Falcon-Barroso et al. find values of the Sersic parameter, n[1], ranging from 0.9 to 6.2 for their sample of bulges, with a median value of 2.9.

The light profiles of the disc and bulge components are described by equations (D.1) and (D.7), respectively. I assume smooth, elliptical isophotes to calculate the semi-major axis, R_{iso}, of the limiting isophote. The isophotal magnitude is then the amount of flux within this radius in the galaxy and is given by the sum (in magnitudes) of the contributions from the two components, which are given by equations (D.6) and (D.13). The assumption of elliptical isophotes is a good approximation for nearly face-on galaxies, or for galaxies whose light is dominated by one or other of the components, but clearly not for edge on galaxies with both a disc and a prominent bulge, since the bulge component is less flattened than the disc. In the next section I compare the luminosity functions derived from the mock surveys with that measured from the

[1]The Sersic (1968) profile takes the form $\mu(r) = \mu_e + 2.5 b_n[(r/r_e)^{1/n} - 1]$. The $R^{1/4}$–profile is a special case of this profile with $n = 4$, while the $n = 1$ case corresponds to an exponential profile.

2dFGRS. For this reason, I select galaxies to have isophotal magnitudes $B_{\rm iso} < 19.71$, which corresponds to $b_{\rm J} < 19.45$, the formal magnitude limit of the 2dFGRS.

4.2.5 Luminosity function

I use the $V_{\rm max}$ approach to estimate the LF from the mock survey. As discussed in section 4.1, this estimator is less then ideal for observational purposes because it relies on the assumption of a uniform galaxy distribution, which is only true on the largest scales. In the mock surveys, however, there is no clustering on scales larger than the size of the simulation boxes, so the assumption is valid and the $V_{\rm max}$ method should give a reliable estimate of the true LF. Observational samples are generally corrected for both redshifting of the waveband (k) and the effects of evolution (e). The k-correction is implicitly included in the simulated sample by using the rest-frame B-band magnitudes of the galaxies. No evolution correction is applied to the model galaxies; the evolution in the bright end of the model LF is small (in the B-band), so I expect this to make only a small difference. A more complete treatment would include this effect.

In Figs. 4.5 and 4.6 I show how different isophotal limits affect the derived LF for two different models. The magnitudes plotted in the figures are *total* magnitudes. The use of isophotal magnitudes, or other types of magnitude, as estimates of the total luminosities of galaxies is also an issue and is expected to contribute to uncertainties in the derived LF (Cross & Driver 2002), but this does not really belong under the heading of selection effects, since it is a question about the accuracy of the parameters measured for galaxies that *are* included in the sample. It is clear from Figs. 4.5 and 4.6 that accounting for selection does have an effect on the predictions of the models. Model K from chapter 3 gives the best fit to the observed scale lengths of disc galaxies and should give the most accurate estimate of the magnitude of the effect of surface brightness selection. However, this model does not have an intrinsically faint steep faint-end slope, as a result of the combined effects of feedback and the star formation threshold, so selection effects need only play a small role in determining the observed faint-end slope. It is more interesting to ask if selection effects can reconcile a steep intrinsic faint-end slope with the observed value. For this reason, I also show the results for a variant of model K with weaker feedback in low-mass haloes (Fig. 4.5b) and for

the basic model of chapter 2, which has no star formation threshold (Fig. 4.6). The steepest faint-end slope is achieved for a model with no star formation threshold and $\alpha_{\rm hot} = 1$. The isophotal limit of the 2dFGRS is 24.50 mag arcsec^{-2} in $b_{\rm J}$, equivalent to 24.76 $B-$mag arcsec^{-2}. At this level the best fit to the observed LF is obtained for models with weak feedback and thus intrinsically steep faint-end slopes; models with shallow intrinsic slopes actually *underpredict* the number of faint galaxies. This result is only qualitative because, as I showed in chapter 3, none of the models gives a perfect fit to the observed sizes of real galaxies. Nevertheless, the effect is clearly important.

The dip in the LF around $M_B = -19.5+5\log_{10} h_0$, which is present in all but one of the LFs in Figs. 4.5 and 4.6 was also evident in the intrinsic LF of model K (Fig. 3.21). Here, it is apparently an artefact of the way in which the mock surveys were constructed, since it also affects the basic model, but the two features are clearly related. They seem to have their origins in the surface-densities (or surface-brightnesses) of the galaxies in question, but further investigation is required to identify the cause.

4.3 Discussion and conclusions

Stellar feedback was introduced in galaxy formation models partly to solve the problem of an overabundance of faint galaxies, leading to a steep faint-end slope, in apparent conflict with the observations of the time. The idea is to suppress star formation in low-mass haloes, thus leaving more gas to form stars in more massive haloes at later times. The motivation for strong stellar feedback is less strong now than it was a decade ago, because the slope of the faint end of the LF has steepened progressively as observations push to deeper isophotal limits. This, in itself, is evidence that surface brightness selection effects have been important in the past, but it is a subject of some controversy whether or not current large galaxy surveys have accounted for most of the galaxies missing from the earlier observations. In chapter 3 I showed how, even without selection effects, star formation thresholds in discs provide an alternative mechanism for suppressing star formation in low-mass haloes, with a similar effect on the faint end of the LF as feedback. The aim in this chapter was to find out if selection effects can, either alone or in conjunction with the two physical mechanisms described above, reconcile

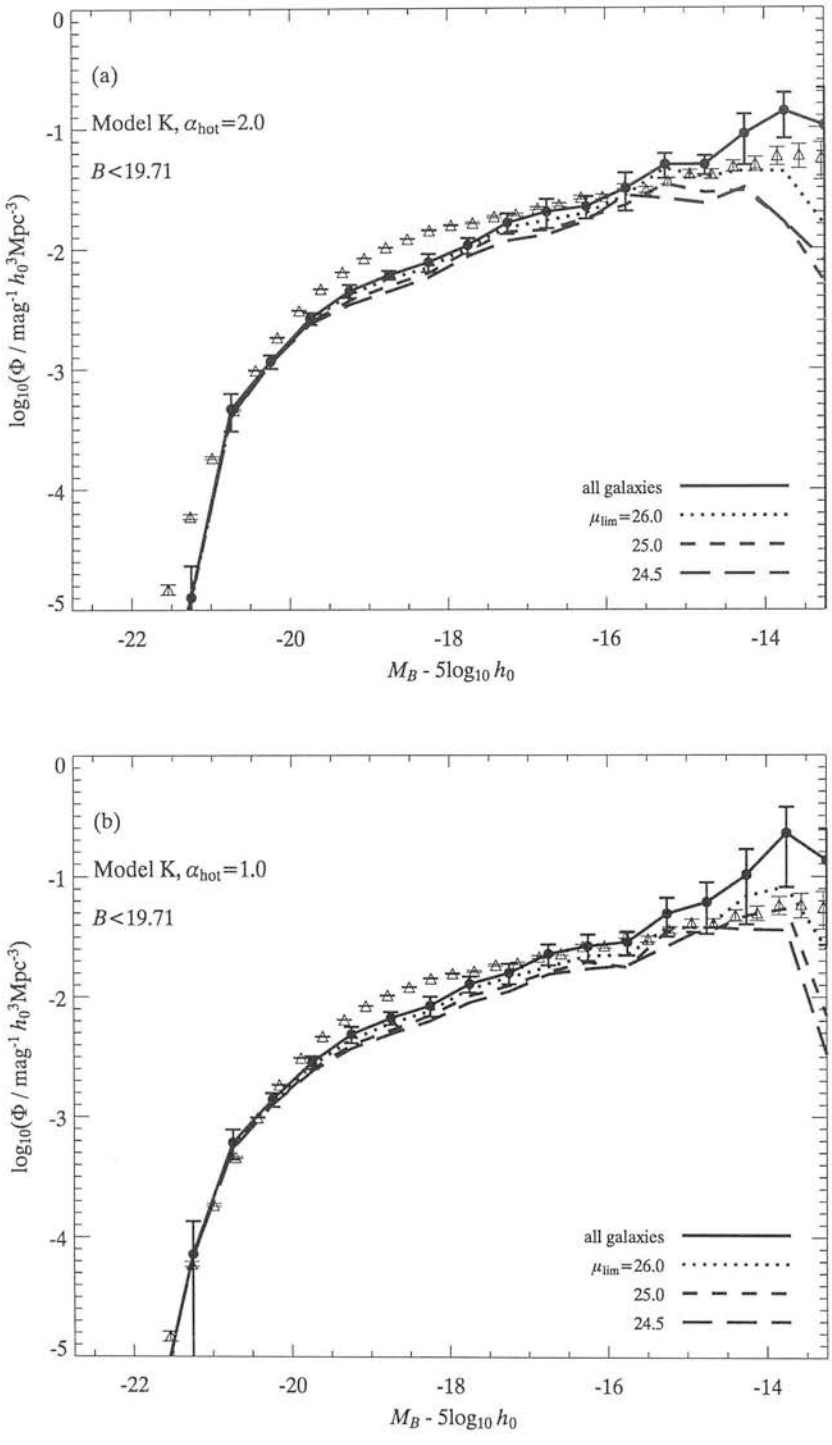

Figure 4.5: B-band luminosity functions (LFs) reconstructed using the maximum-volume estimator from a 125 deg^2 survey with a magnitude limit of $B = 19.71$. The LFs in (a) are for model K with the fiducial feedback slope, $\alpha_{hot} = 2$, and those in (b) are for a variant of the same model with $\alpha_{hot} = 1$. The heavy, solid curves are the intrinsic LFs of the galaxies in the survey volume, while the remaining curves are for magnitude-limited samples with different limiting isophotes. The triangles with error bars give the LF of the 2dFGRS (Norberg et al. 2002).

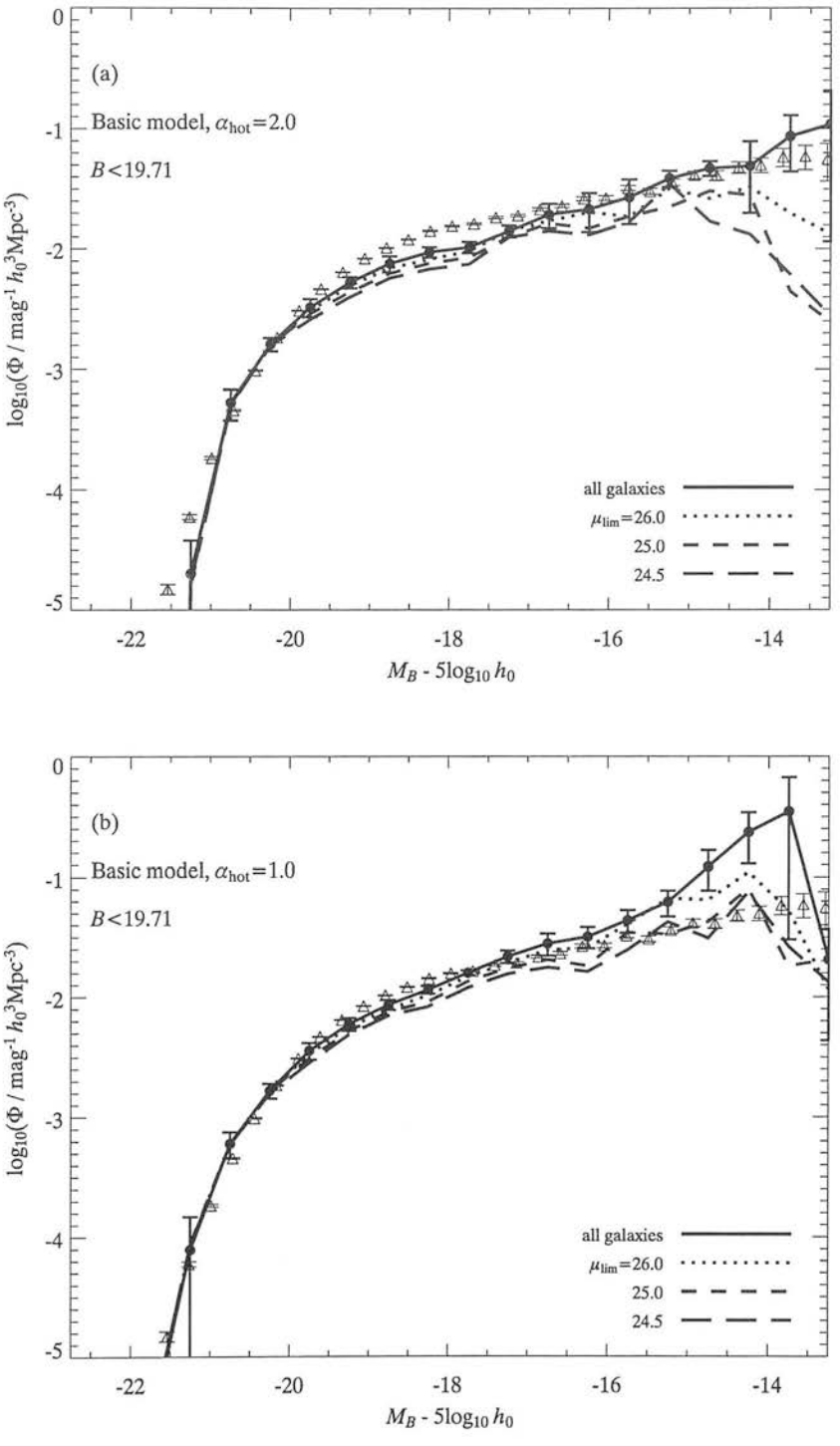

Figure 4.6: B-band luminosity functions (LFs) reconstructed using the maximum-volume estimator from a 125 deg^2 survey with a magnitude limit of $B = 19.71$. The LFs in (a) are for the basic model with the fiducial feedback slope, $\alpha_{\rm hot} = 2$, and those in (b) are for a variant of the same model with $\alpha_{\rm hot} = 1$. The heavy, solid curves are the intrinsic LFs of the galaxies in the survey volume, while the remaining curves are for magnitude-limited samples with different limiting isophotes. The triangles with error bars give the LF of the 2dFGRS (Norberg et al. 2002).

the predictions of galaxy formation models with the observed faint-end slope of the LF. Although the results cannot be considered conclusive, because of the outstanding problem of matching the observed bivariate brightness distribution, my investigation shows that this is indeed a possibility: by selecting model galaxies from a mock survey on the basis of their isophotal magnitudes, it is possible to achieve agreement with the observed LF even for models with intrinsically steep faint-end slopes.

This result removes one of the motivations for the inclusion of strong stellar feedback in galaxy formation models. Feedback may still be neccessary for other reasons, particularly to prevent catastrophic cooling into small structures at early times, leading to an 'angular momentum catastrophe'. However, observational evidence, in the form of strong supernova-driven winds, for the importance of feedback in normal galaxies at the present-day has yet to be found. By 'normal', I mean galaxies that are not undergoing a strong burst of star formation. Starbursts of this kind are usually concentrated in the nuclei of galaxies and are often often associated with strong stellar winds. It has been suggested that the efficiency of feedback is dependant on the size of the star forming region, in which case it might be expected that feedback was important early in the universe, when starbursts were more common, but that it is of less importance in present-day galaxies.

The results presented in this chapter are interesting in their own right. Of equal importance is the method used to obtain them. The approach of constructing mock galaxy surveys from cosmological N-body simulations, by combining them with a phenomenological recipe for galaxy formation, is likely to have many practical applications, not only to the interpretation of existing observational samples, but also in planning future observations.

Chapter 5

Extending the model

The simple description of galaxy discs set out in chapters 2 and 3 provides a satisfatory explanation of several key observations, but the supposedly most realistic model is still unable to match the observed sizes and circular velocities of spiral galaxies, producing discs that are systematically too small and rotate too rapidly. Preventing overmerging by preserving substructure within haloes no longer seems to solve the problem, as it did in the much simpler model of vKJP, so it is necessary to look to other parts of the model for a solution. A promising possibility is that galaxy haloes do not have cuspy profiles, as suggested by dissipationless N-body simulations, but instead have large, constant density cores, as observed in low surface-brightness dwarf galaxies. In this chapter, I show, briefly, how I include the possibility of different halo profiles in the model and discuss a number of other possible improvements. The ingredients presented here do not constitute a complete or thoroughly tested model and the results are only preliminary. I include them to give an idea of the potential for future investigations.

5.1 Self-consistent disc profiles and rotation curves

The dynamics of the disc are governed by a gravitational potential with three components: due to the disc itself, the bulge (if one exists) and the dark matter halo. The potentials add linearly so that the circular velocity is a sum in quadrature of the

contributions of the three components:

$$V^2_{\rm disc}(r) = V^2_{\rm c,\,disc}(r) + V^2_{\rm c,\,bulge}(r) + V^2_{\rm c,\,halo}(r). \tag{5.1}$$

The contribution made to the rotation curve by each component is described below. I assume that the disc retains its exponential mass profile at all times and ignore the response of the bulge and the halo to each other and to the disc. The effect of self-gravity is thus simply to reduce the exponential scale length of the disc. As an initial guess I take the value of the scale length for a disc with a flat rotation curve (equation 2.36). I then follow MMW in solving iteratively to find a mutually-consistent disc profile and rotation curve. Briefly, the procedure consists of

1. using equation 5.1 to calculate the rotation curve of the disc for the initial scale length;

2. choosing a new scale length, such that the angular momentum of the disc,

$$j_{\rm disc} = \frac{R_{\rm d}}{8} \int_0^\infty V_{\rm c}(y) y^2 {\rm e}^{-2y} {\rm d}y, \tag{5.2}$$

   is conserved; and then

3. re-calculating the rotation curve for the new scale length, taking into account the fact that the disc is now sampling a different part of the halo rotation curve.

Steps 2 and 3 are repeated until the values obtained for the scale length converge to within 1 part in 1000, which typically occurs after a small number ($\lesssim 10$) of iterations.

Halo component

In addition to the singular isothermal sphere, I consider two alternative profiles. The 'universal' NFW profile, given by equation (B.5), is found to give a good fit to the density profiles of haloes in N-body simulations. However, the rotation curves of low surface-brightness dwarf galaxies are better fit by profiles with constant density cores (de Blok & McGaugh 1997; Swaters, Madore & Trewhella 2000; de Blok & Bosma 2002); an example of such a profile is the pseudo-isothermal sphere (equation B.10). The contributions of these two profiles to the rotation curve of the disc are given by equations (B.7) and (B.12), respectively

Disc component

The rotation curve of an isolated, self-gravitating, exponential disc is given by Freeman (1970):

$$V^2_{\mathrm{c,disc}}(y) = \frac{2GM_{\mathrm{disc}}}{R_{\mathrm{d}}} y^2 \left[I_0(y)K_0(y) - I_1(y)K_1(y)\right], \tag{5.3}$$

where M_{disc} is the total mass (stars+gas) of the disc, $I_i(y)$ and $K_i(y)$ are modified Bessel functions and $y = R/2R_{\mathrm{d}}$.

Bulge component

The contribution of the stellar spheroidal component can be significant for the rotation curves of small discs. It can have a stabilising effect on a disc that would otherwise be gravitationally unstable, although I do not consider this possibility here. Ellipticals and the bulges of early-type spirals have light profiles that are, in projection, described by an $R^{1/4}$–law (equation D.7). I assume that the spheroid light traces the mass; this is more likely to be the case for spheroids than it is for discs because, observationally, spheroids have uniformly old stellar populations and are generally gas-poor. A 3-dimensional density profile that, projected onto a plane, closely resembles the $R^{1/4}$–law profile is

$$\rho(r) = \frac{3M_{\mathrm{bulge}}}{8\pi} \frac{a_{\mathrm{b}}}{r^{3/2}(r + a_{\mathrm{b}})^{5/2}} \tag{5.4}$$

(Dehnen 1993). A bulge described by equation (5.4) makes a contribution to the circular velocity of

$$V^2_{\mathrm{c,bulge}}(w) = \frac{GM_{\mathrm{bulge}}}{a_{\mathrm{b}}} \frac{w^{1/2}}{(w+1)^{3/2}}, \tag{5.5}$$

where $w = r/a_{\mathrm{b}}$. The scale radius, a_{b}, is related to the 2-dimensional half-light radius, R_{e}, of the $R^{1/4}$–law by $R_{\mathrm{e}} = 1.28a_{\mathrm{b}}$ (Binney & Merrifield 1998, table 4.5).

The virial criterion for a self-gravitating system described by equation (5.4) can be written in the form

$$r_{\mathrm{b}} \simeq 0.4 \frac{GM_{\mathrm{bulge}}}{\langle v^2 \rangle_{\mathrm{bulge}}}, \tag{5.6}$$

where r_{b} is the half-mass radius (this can be checked by direct integration of the expression for the binding energy). The half-mass radius is related to the scale radius in equation (5.4) by $a_{\mathrm{b}} = 0.5874r_{\mathrm{b}}$. If the spheroid is indeed a self-gravitating stellar system then equation (5.6) can be used to predict the half-mass radius given the stellar

velocity dispersion. This latter quantity is not predicted by my simple model, although any model for bulge formation should be able to predict at least this basic property. Instead, as in chapter 4, I use the velocity dispersion of the halo as a first estimate.

If self-gravity is important, or if the halo profile is not isothermal, then the disc rotation curve will not in general be flat. For a general rotation curve, $V_{\rm disc}(R)$, the critical density no longer takes the simple form of equation (3.6). Instead we have to use the general expression (equation 3.5), substituting in for the epicyclic frequency κ, given by equation (3.4), at each radius. Thus, the exact radial dependence of $\Sigma_{\rm crit}$ depends on the rotation curve of the galaxy. However, for most realistic rotation curves, the critical density behaves in the same qualitative way as in the simple case outlined in chapter 3, that is, it decreases more steeply than the exponential surface density profile at small radii and is shallower at large radii. Hence, it is still possible to define the two critical radii, $R_{\rm crit}^{(1)}$ and $R_{\rm crit}^{(2)}$, as the minimum and maximum radii for which $\Sigma_{\rm gas} \geqslant \Sigma_{\rm crit}$. Fig. 5.1 shows the situation for four typical galaxies: two that have widespread massive star formation, one in which the gas density is barely above the threshold at any radius and a fourth that is completely sub-threshold. Note that in all four cases the rotation curve is approximately flat at radii greater than 1–2 kpc.

The effect of different halo profiles on the Tully–Fisher relation is shown in Fig. 5.2. Rather disappointingly, a constant density core, even quite a large one ($r_{\rm c} = r_{\rm vir}/10$), appears to do little to change the circular velocities of the more massive discs. At first sight, it would appear that the halo profile is not of primary importance to the Tully–Fisher relation. However, there are still some outstanding issues, particularly with the way that the virial velocity is calculated from the N-body simulations when the profiles considered would have different binding energies to that assumed in applying the virial test (see the discussion at the end of section 3.2.3).

5.2 Dust in discs

Now that we have basic structural parameters for the model galaxies, it should be possible to apply a more realistic model for dust extinction which takes account of the geometries of galaxies. A simple model (but one that is more realistic than the scaling

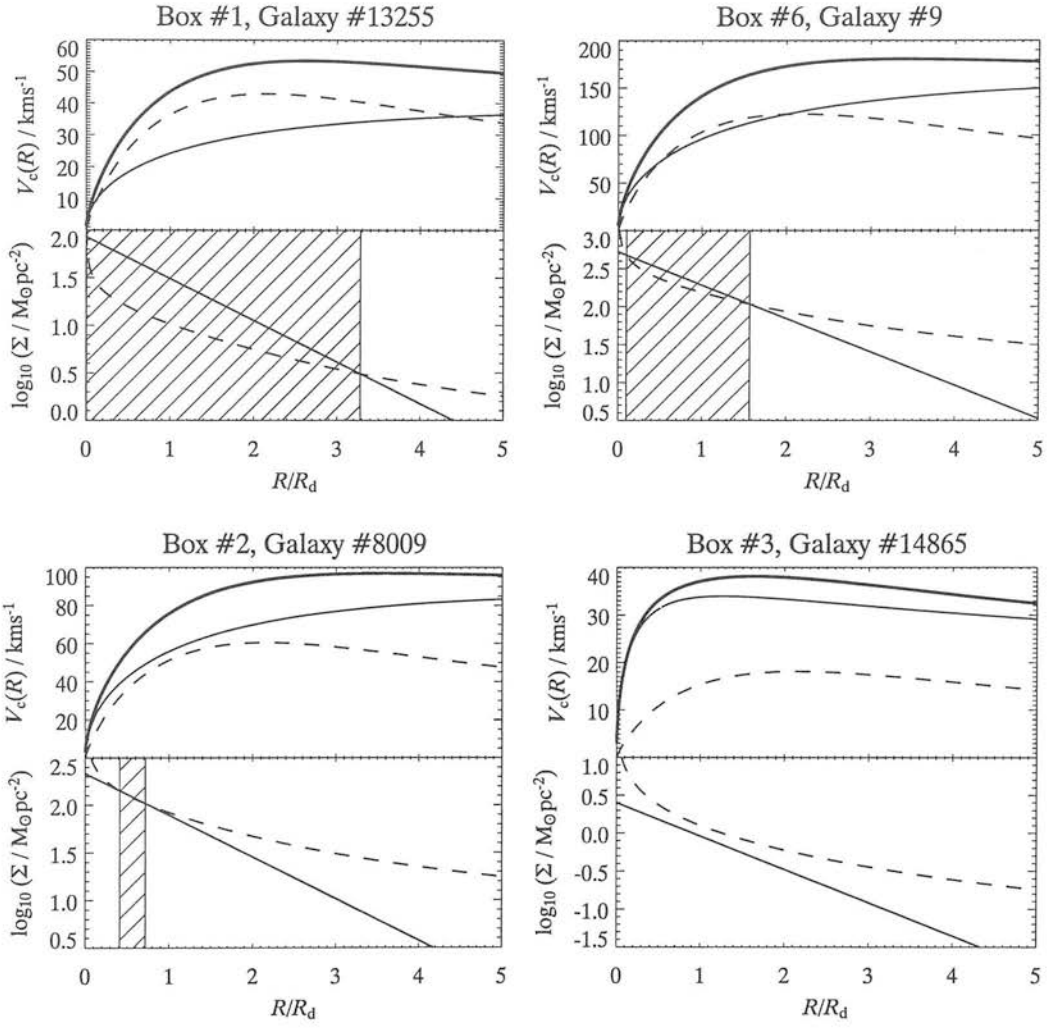

Figure 5.1: The star formation threshold in four typical galaxies with realistic rotation curves. In the top half of each panel I plot the rotation curve of the disc (heavy, solid line) separated into components due to the halo (light, solid line) and the disc itself (dashed line). In the bottom half I plot the critical density (dashed line) calculated according to equation (3.5) and the density profile of the gas disc (solid line). The hatched area is the star forming region, which I identify with the optical disc of the galaxy.

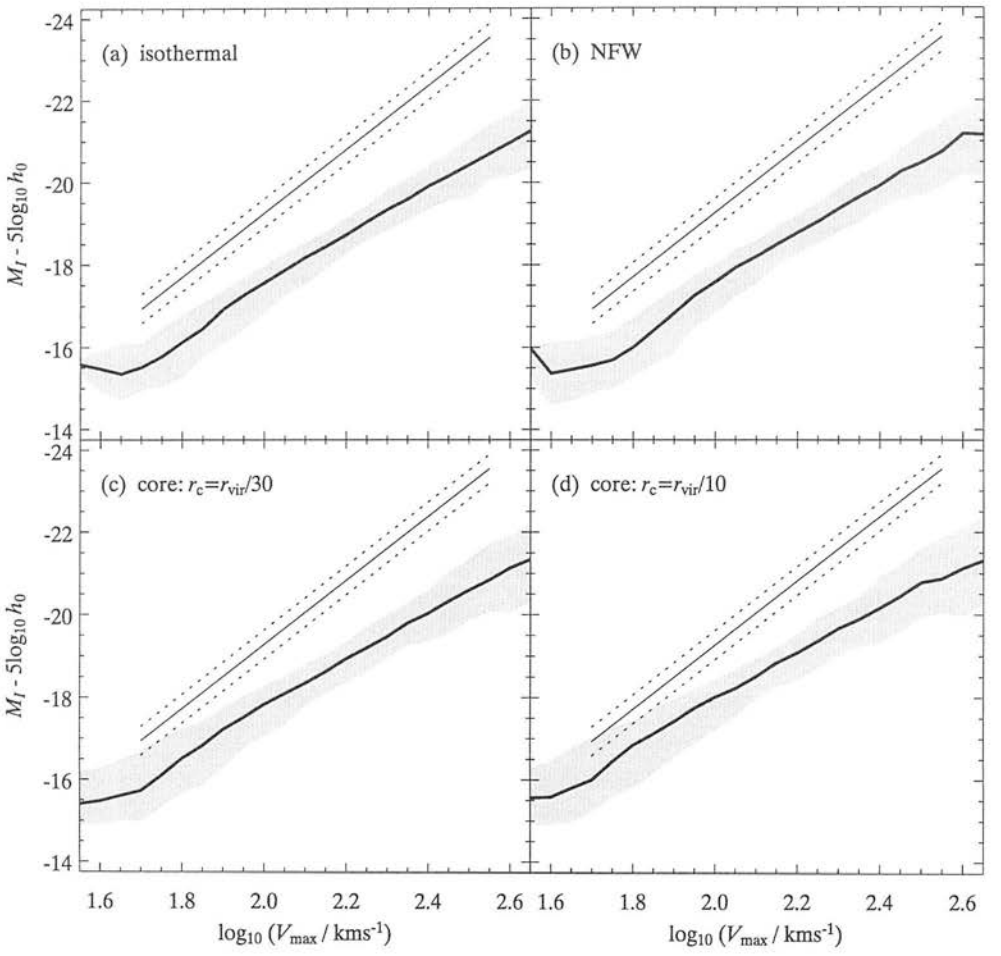

Figure 5.2: The I-band Tully–Fisher relation (TFR) for different halo profiles: (a) a singular isothermal sphere, (b) the NFW profile, (c) a pseudo-isothermal sphere with core radius $r_c = r_{vir}/30$ and (d) another pseudo-isothermal sphere with $r_c = r_{vir}/10$. The heavy solid line is the median relation and the shaded area marks the 10% and 90% quartiles of the distribution. The remaining lines mark the mean observed TFR (see Fig. 2.22 for details)

law adopted previously) is one in which the amount of extinction scales with the surface density of cold gas in the disc and with its metallicity. Guiderdoni & Rocca-Volmerange (1987) estimate the face-on optical depth of a disc galaxy as

$$\tau_\lambda^Z = \left(\frac{A_\lambda}{A_V}\right)_{Z_\odot} \left(\frac{Z}{Z_\odot}\right)^s \frac{\langle N_H \rangle}{N_H^0}, \tag{5.7}$$

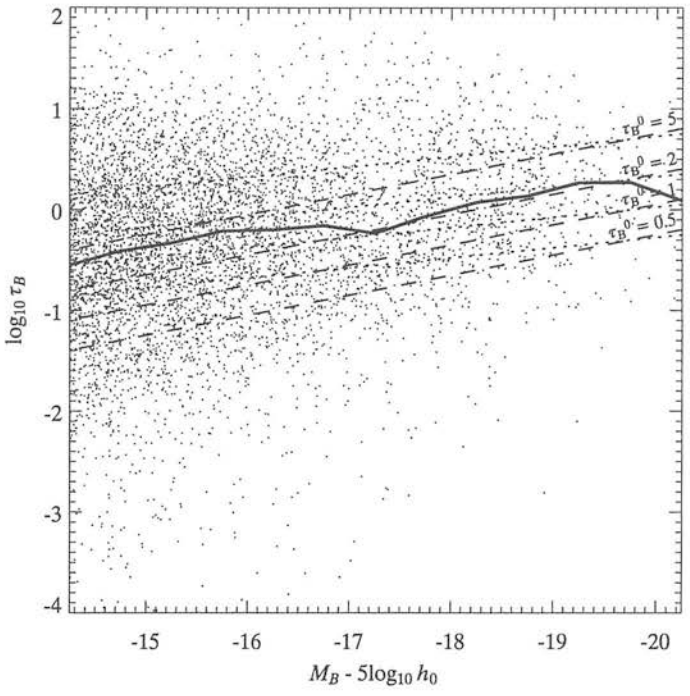

Figure 5.3: The optical depths of individual galaxies as a function of absolute B-band magnitude. The heavy solid curve marks the median value of $\log_{10}\tau_B$. Diagonal lines mark the predictions of the scaling law of Wang & Heckman (1996) for different values of τ_B^0 and for slopes of $\beta = 0.5$ (dashed lines) and $\beta = 0.25$ (dotted lines). Lines for the same value of τ_B^0 agree at $M_B \simeq -19.7 + 5\log_{10} h_0$.

where $\langle N_{\rm H}\rangle$ is the average column depth of hydrogen over the disc. For $(A_\lambda/A_V)_{Z_\odot}$ I take the mean Galactic extinction curve of Cardelli, Clayton & Mathis (1989). Guiderdoni & Rocca-Volmerange (1987) find a scaling with metallicity of $s = 1.6$ for $\lambda > 2000$ A, which covers the spectral region I am interested in. As a measure of the 'average' hydrogen column depth, I take the surface density at the half-light radius, related to the gas surface density, $\Sigma_{\rm g}(r_{\rm half})$, by

$$\langle N_{\rm H}\rangle = \frac{\Sigma_{\rm g}(r_{\rm half})}{1.4 m_{\rm h}} \tag{5.8}$$

In Fig. 5.3 I plot the optical depth predicted by the above model as a function of B-band magnitude. I also show the predictions of the scaling law of Wang & Heckman (1996). The average extinction agrees well with the scaling law parameterisation, but there is a large degree of scatter which reflects variations in the sizes, metallicities and

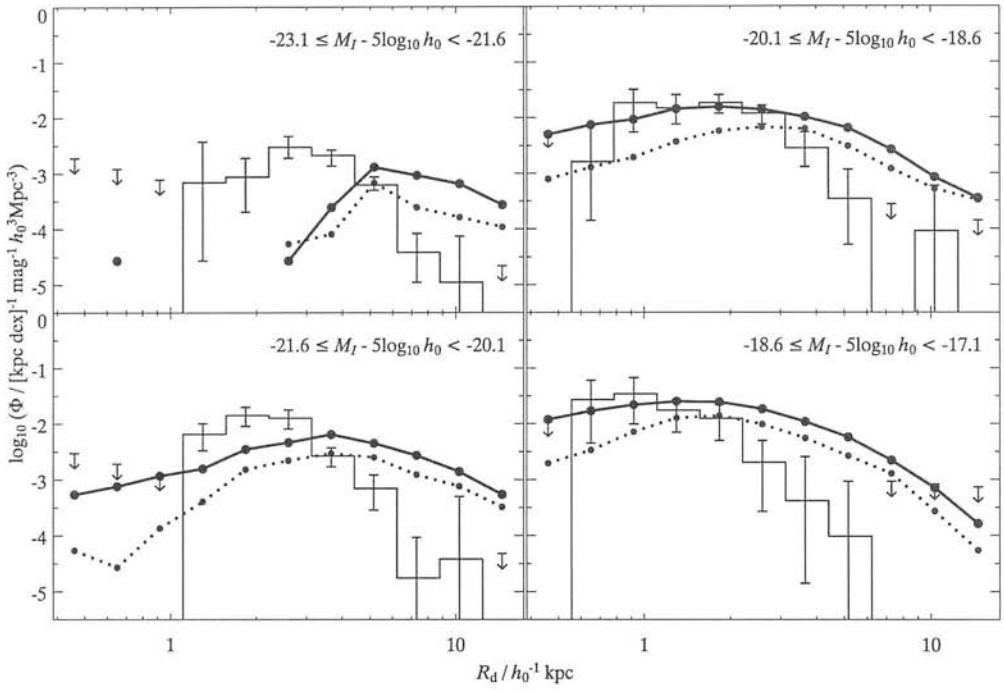

Figure 5.4: The effect on the bivariate distribution of adding a star formation threshold, when the amount of dust extinction scales with the gas surface density. The solid curve is for a model with a threshold. The solid histogram with error bars is the distribution of de Jong & Lacey (2000).

gas contents of galaxies of a given luminosity.

The dependence of the optical depth on the surface density of the gas disc has interesting consequences for the distribution of scale lengths at a given luminosity, as illustrated in Fig. 5.4. The existence of star formation thresholds now has more of an effect on small scale-length galaxies, which suffer more extinction for a given mass of cold gas. This may be one way of explaining away the excess of small scale-length galaxies seen in the simulations — a possibility that is certainly worthy of further investigation.

5.3 Future directions

I have presented what I hope is a simple, yet realistic, method for modelling the formation of galaxies in haloes identified in N-body simulations. One of the main motivations

for the work presented in this thesis was the need for a model that can predict the morphologies of galaxies, in particular the gross distinction between discs and bulges. This is desirable for several reasons:

1. to increase the predictive power of the model to encompass such properties as disc sizes, bulge-to-disc ratios, etc. and their variation with epoch and environment, and in doing so to impose additional constraints on the model;

2. to make the model more realistic and its predictions more accurate, for example enabling the rotational velocity of the disc, including the effects of its own self-gravity, to be used in the Tully–Fisher relation, rather than the virial velocity of the halo;

3. to enable a fairer comparison with observations, which are often biased by morphological or surface-brightness dependent selection effects, or a combination of the two.

There are, of course, always improvements that can be made to the model in order to make it even more realistic. An outstanding problem is explaining the zero-point of the observed Tully–Fisher relation. Preventing overmerging no longer seems to help, as it did in the simpler model of vKJP, so we have to look to other aspects of the model, such as the shape of the halo density profile or the details of the disc formation process. The formation and evolution of galaxy discs in particular has been studied in detail by several authors (e.g. Dalcanton, Spergel & Summers 1997; MMW; Firmani & Avila-Reese 2000; van den Bosch 2000, 2001; Ferguson & Clarke 2001). Models based on N-body simulations give no particular advantage over analytic and semi-analytic models in this area. The last of the papers listed above address another key issue for galaxy formation models: that of the origin of the exponential profile, and the questions of whether the exponential light profile reflects the underlying mass profile and whether this in turn reflects the infall profile or is the result of subsequent evolution. The exponential disc profile is one of the fundamental assumptions of the model presented in this thesis. If the assumption turned out to be incorrect, it could have important consequences, particularly on the star formation threshold. Ferguson & Clarke restate

the idea, originally proposed by Lin & Pringle (1987), of viscous redistribution of angular momentum in galaxy discs, which leads naturally to an exponential stellar profile, for initial gas profiles that are, in fact, much flatter. This is an interesting possibility that could be incorporated into the model without too much difficulty. However, it is not clear that building an even more detailed recipe for galaxy discs into the model adds anything to our understanding of what is going on, especially given the uncertainties in some of the more fundamental aspects of the model, such as the definition and properties of the haloes themselves. Again, this task is better-suited to analytic and semi-analytic models such as those listed above. That is not to say that incorporating some of the *results* of more detailed studies into a phenomenological model such as this, when it is clear that they will make a difference to the predictions of the model, is not a useful excercise. An example of how this can be useful is the use of the approximate expressions of MMW for the scale lengths and circular velocities self-gravitating discs in haloes with realistic profiles. This clearly made the model more realistic; the fact that it resulted in a poorer match to the observational data is an indication of a more fundamental problem with the model. However, overall, the model should be treated as a link between the more detailed schemes mentioned above and the overall cosmological context, rather than a complete description of galaxy formation.

In the future, as observations improve and as our understanding of galaxy formation advances, it may be both desirable and practical to build models that contain both cosmological scales and detailed descriptions of individual galaxies. Currently, however, there are more fundamental questions to be answered: *How important is feedback in galaxy formation? What is the mechanism responsible for preventing cooling flows in massive clusters? What role do active galactic nuclei (AGN) play in galaxy formation.* Such global questions make detailed modelling of one particular aspect of galaxy formation an unneccessary extravagance. This is not a negative conclusion; on the contrary, the real strength of this type of modelling lies not in detailed solutions to single aspects of galaxy formation but in predicting simple, global properties for a cosmologically significant number of galaxies.

List of Figures

List of Tables

Bibliography

Barnes J., Hut P., 1986, Nature, 324, 446

Barnes J. E., 1988, ApJ, 331, 699

Barnes J. E., 1992, ApJ, 393, 484

Barnes J. E., Hernquist L., 1996, ApJ, 471, 115

Baugh C. M., Cole S., Frenk C. S., 1996, MNRAS, 283, 1361

Baugh C. M., Cole S., Frenk C. S., Lacey C. G., 1998, ApJ, 498, 504

Bertschinger E., Gelb J. M., 1991, Computers in Physics, 5, 164

Binney J., Merrifield S., 1998, *Galactic astronomy*, Princeton: Princeton University Press

Binney J., Tremaine S., 1987, *Galactic dynamics*, Princeton: Princeton University Press

Blair M., Gilmore G., 1982, PASP, 94, 742

Blanton M. R., et al., 2001, AJ, 121, 2358

Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440

Boselli A., Lequeux J., Gavazzi G., 2002, A&A, 384, 33

Bothun G. D., Impey C. D., Malin D. F., 1991, ApJ, 376, 404

Bothun G. D., Impey C. D., Malin D. F., Mould J. R., 1987, AJ, 94, 23

Bower R. G., 1991, MNRAS, 248, 332

Brainerd T. G., Oaxaca Wright C., Goldberg D. M., Villumsen J. V., 1999, ApJ, 524, 9

Bruzual A. G., Charlot S., 1993, ApJ, 405, 538

Burkholder V., Impey C., Sprayberry D., 2001, AJ, 122, 2318

Burles S., Tytler D., 1998, ApJ, 499, 699

Butcher H., Oemler A., 1978, ApJ, 226, 559

Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245

Chandrasekhar S., 1943, ApJ, 97, 255

Charlot S., Worthey G., Bressan A., 1996, ApJ, 457, 625

Cole S., Aragon-Salamanca A., Frenk C. S., Navarro J. F., Zepf S. E., 1994, MNRAS, 271, 781

Cole S., et al., 2001, MNRAS, 326, 255

Cole S., Lacey C., 1996, MNRAS, 281, 716

Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRAS, 319, 168

Cole S., Weinberg D. H., Frenk C. S., Ratra B., 1997, MNRAS, 289, 37

Combes F., Debbasch F., Friedli D., Pfenniger D., 1990, A&A, 233, 82

Courteau S., de Jong R. S., Broeils A. H., 1996, ApJ, 457, L73

Cross N., Driver S. P., 2002, MNRAS, 329, 579

Cross N., et al., 2001, MNRAS, 324, 825

Dalcanton J. J., Spergel D. N., Summers F. J., 1997, ApJ, 482, 659

Davies J. I., 1990, MNRAS, 244, 8

Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371

de Blok W. J. G., Bosma A., 2002, A&A, 385, 816

de Blok W. J. G., McGaugh S. S., 1997, MNRAS, 290, 533

de Jong R. S., 1996, A&A, 313, 45

de Jong R. S., Lacey C., 2000, ApJ, 545, 781

de Vaucouleurs G., 1948, Annales d'Astrophysique, 11, 247

Dehnen W., 1993, MNRAS, 265, 250

Disney M. J., 1976, Nature, 263, 573

Dopita M. A., Ryder S. D., 1994, ApJ, 430, 163

Dressler A., Oemler A. J., Butcher H. R., Gunn J. E., 1994, ApJ, 430, 107

Efstathiou G., Ellis R. S., Peterson B. A., 1988, MNRAS, 232, 431

Eggen O. J., Lynden-Bell D., Sandage A. R., 1962, ApJ, 136, 748

Eke V. R., Cole S., Frenk C. S., 1996, MNRAS, 282, 263

Evrard A. E., 1991, MNRAS, 248, 8P

Ewald P. P., 1921, Ann. Physik, 64, 253

Fabian A. C., 1994, ARA&A, 32, 277

Falcón-Barroso J., Peletier R. F., Balcells M., 2002, MNRAS, 335, 741

Fall S. M., 1979, Nature, 281, 200

Fall S. M., Efstathiou G., 1980, MNRAS, 193, 189

Ferguson A. M. N., Clarke C. J., 2001, MNRAS, 325, 781

Ferguson H. C., Sandage A., 1988, AJ, 96, 1520

Firmani C., Avila-Reese V., 2000, MNRAS, 315, 457

Firmani C., D'Onghia E., Chincarini G., Hernández X., Avila-Reese V., 2001, MNRAS, 321, 713

Folkes S., et al., 1999, MNRAS, 308, 459

Freedman W. L., et al., 2001, ApJ, 553, 47

Freeman K. C., 1970, ApJ, 160, 811

Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, AJ, 111, 1748

Gardner J. P., Sharples R. M., Frenk C. S., Carrasco B. E., 1997, ApJ, 480, L99

Ghigna S., Moore B., Governato F., Lake G., Quinn T., Stadel J., 2000, ApJ, 544, 616

Gingold R. A., Monaghan J. J., 1977, MNRAS, 181, 375

Giovanelli R., Haynes M. P., da Costa L. N., Freudling W., Salzer J. J., Wegner G., 1997, ApJ, 477, L1

Governato F., Moore B., Cen R., Stadel J., Lake G., Quinn T., 1997, New Astronomy, 2, 91

Graham A. W., 2001, AJ, 121, 820

Guiderdoni B., Rocca-Volmerange B., 1987, A&A, 186, 1

Helly J. C., Cole S., Frenk C. S., Baugh C. M., Benson A., Lacey C., 2003, MNRAS, 338, 903

Hernquist L., 1992, ApJ, 400, 460

Hernquist L., 1993, ApJ, 409, 548

Hernquist L., Bouchet F. R., Suto Y., 1991, ApJS, 75, 231

Hernquist L., Spergel D. N., Heyl J. S., 1993, ApJ, 416, 415

Heyl J. S., Hernquist L., Spergel D. N., 1994, ApJ, 427, 165

Hoffman Y., Ribak E., 1991, ApJ, 380, L5

Hoyle F., 1953, ApJ, 118, 513

Huang J.-S., Glazebrook K., Cowie L. L., Tinney C., 2003, ApJ, 584, 203

Huchtmeier W. K., Richter O.-G., 1988, A&A, 203, 237

Impey C., Bothun G., Malin D., 1988, ApJ, 330, 634

Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, MNRAS, 321, 372

Jimenez R., Padoan P., Matteucci F., Heavens A. F., 1998, MNRAS, 299, 123

Katz N., Gunn J. E., 1991, ApJ, 377, 365

Kauffmann G., 1996, MNRAS, 281, 475

Kauffmann G., Charlot S., 1998, MNRAS, 294, 705

Kauffmann G., Colberg J. M., Diaferio A., White S. D. M., 1999, MNRAS, 303, 188

Kauffmann G., White S. D. M., 1993, MNRAS, 261, 921

Kauffmann G., White S. D. M., Guiderdoni B., 1993, MNRAS, 264, 201

Kennicutt R. C., 1989, ApJ, 344, 685

Kennicutt R. C., 1998, ApJ, 498, 541

Kennicutt R. C., Tamblyn P., Congdon C. E., 1994, ApJ, 435, 22

Kochanek C. S., et al., 2001, ApJ, 560, 566

Komatsu E., et al., 2003, ApJ, submitted (astro-ph/0302223)

Krauss L. M., 2003, ESO-CERN-ESA Symposium on Astronomy, Cosmology and Fundamental Physics, preprint (astro-ph/0301012)

Kregel M., van der Kruit P. C., de Grijs R., 2002, MNRAS, 334, 646

Lacey C., Cole S., 1994, MNRAS, 271, 676

Lahav O., et al., 2002, MNRAS, 333, 961

Lemson G., Kauffmann G., 1999, MNRAS, 302, 111

Lin D. N. C., Pringle J. E., 1987, ApJ, 320, L87

Loveday J., 1996, MNRAS, 278, 1025

Loveday J., 2000, MNRAS, 312, 557

Loveday J., Peterson B. A., Efstathiou G., Maddox S. J., 1992, ApJ, 390, 338

Lynden-Bell D., 1967, MNRAS, 136, 101

Maddox S. J., Sutherland W. J., Efstathiou G., Loveday J., Peterson B. A., 1990, MNRAS, 247, 1P

Madgwick D. S., et al., 2002, MNRAS, 333, 133

Marigo P., 2001, A&A, 370, 194

Martin C. L., Kennicutt R. C., 2001, ApJ, 555, 301

Marzke R. O., da Costa L. N., Pellegrini P. S., Willmer C. N. A., Geller M. J., 1998, ApJ, 503, 617

Mathewson D. S., Ford V. L., 1996, ApJS, 107, 97

Mathewson D. S., Ford V. L., Buchhorn M., 1992, ApJS, 81, 413

McGaugh S. S., Bothun G. D., Schombert J. M., 1995, AJ, 110, 573

Metcalfe N., Shanks T., Fong R., Jones L. R., 1991, MNRAS, 249, 498

Mihos J. C., Hernquist L., 1994a, ApJ, 425, L13

Mihos J. C., Hernquist L., 1994b, ApJ, 431, L9

Mihos J. C., Hernquist L., 1996, ApJ, 464, 641

Miller G. E., Scalo J. M., 1979, ApJS, 41, 513

Miralda-Escude J., Babul A., 1995, ApJ, 449, 18

Mo H. J., Mao S., 2000, MNRAS, 318, 163

Mo H. J., Mao S., White S. D. M., 1998, MNRAS, 295, 319

Moore B., Ghigna S., Governato F., Lake G., Quinn T., Stadel J., Tozzi P., 1999, ApJ, 524, L19

Moore B., Lake G., Katz N., 1998, ApJ, 495, 139

Mould J. R., Akeson R. L., Bothun G. D., Han M., Huchra J. P., Roth J., Schommer R. A., 1993, ApJ, 409, 14

Naab T., Burkert A., 2001a, in J. G. Funes, S. J. Corsini, & E. M. Corsini (eds.), *Galaxy Disks and Disk Galaxies*, ASP Conf. Ser., 230, San Francisco: Astronomical Society of the Pacific, pp 453–454

Naab T., Burkert A., 2001b, in J. G. Funes, S. J. Corsini, & E. M. Corsini (eds.), *Galaxy Disks and Disk Galaxies*, ASP Conf. Ser., 230, San Francisco: Astronomical Society of the Pacific, pp 451–452

Naab T., Burkert A., Hernquist L., 1999, ApJ, 523, L133

Navarro J. F., Frenk C. S., White S. D. M., 1995, MNRAS, 275, 56

Navarro J. F., Frenk C. S., White S. D. M., 1997, ApJ, 490, 493

Norberg P., et al., 2002, MNRAS, 332, 827

O'Neil K., Bothun G., 2000, ApJ, 529, 811

O'Neil K., Bothun G. D., Cornell M. E., 1997, AJ, 113, 1212

O'Neil K., Bothun G. D., Schombert J., 2000, AJ, 119, 136

Peacock J. A., 1999, *Cosmological physics*, Cambridge: Cambridge University Press

Peacock J. A., Smith R. E., 2000, MNRAS, 318, 1144

Peebles P. J. E., 1969, ApJ, 155, 393

Peebles P. J. E., 1971, A&A, 11, 377

Peletier R. F., Balcells M., 1996, AJ, 111, 2238

Perlmutter S., et al., 1997, ApJ, 483, 565

Perlmutter S., et al., 1999, ApJ, 517, 565

Phillipps S., Disney M. J., Kibblewhite E. J., Cawson M. G. M., 1987, MNRAS, 229, 505

Pierce M. J., Tully R. B., 1992, ApJ, 387, 47

Plummer H. C., 1911, MNRAS, 71, 460

Press W. H., Schechter P., 1974, ApJ, 187, 425

Sadler E. M., Oosterloo T. A., Morganti R., Karakas A., 2000, AJ, 119, 1180

Sage L. J., 1993, A&A, 272, 123

Salpeter E. E., 1955, ApJ, 121, 161

Sandage A., Tammann G. A., Yahil A., 1979, ApJ, 232, 352

Scalo J. M., 1986, Fundamentals of Cosmic Physics, 11, 1

Schechter P., 1976, ApJ, 203, 297

Schmidt M., 1959, ApJ, 129, 243

Schombert J. M., Bothun G. D., Impey C. D., Mundy L. G., 1990, AJ, 100, 1523

Schombert J. M., Bothun G. D., Schneider S. E., McGaugh S. S., 1992, AJ, 103, 1107

Sersic J. L., 1968, *Atlas de galaxias australes*, Cordoba: Observatorio Astronomico

Sheth R. K., Tormen G., 2002, MNRAS, 329, 61

Silk J., 1997, ApJ, 481, 703

Simien F., de Vaucouleurs G., 1986, ApJ, 302, 564

Sleath J. P., Alexander P., 1995, MNRAS, 275, 507

Somerville R. S., 1997, Ph.D. Thesis

Somerville R. S., Primack J. R., 1999, MNRAS, 310, 1087

Spergel D. N., et al., 2003, ApJ, submitted (astro-ph/0302209)

Spitzer L. J., 1969, ApJ, 158, L139+

Sprayberry D., 1994, Ph.D. Thesis

Sprayberry D., Impey C. D., Irwin M. J., Bothun G. D., 1997, ApJ, 482, 104

Sutherland R. S., Dopita M. A., 1993, ApJS, 88, 253

Swaters R. A., Madore B. F., Trewhella M., 2000, ApJ, 531, L107

Tegmark M., 1996, ApJ, 464, L35

The L. S., White S. D. M., 1986, AJ, 92, 1248

Toomre A., 1964, ApJ, 139, 1217

Tully R. B., Fisher J. R., 1977, A&A, 54, 661

Tully R. B., Pierce M. J., Huang J., Saunders W., Verheijen M. A. W., Witchalls P. L., 1998, AJ, 115, 2264

Tytler D., Fan X.-M., Burles S., 1996, Nature, 381, 207

van de Weygaert R., Bertschinger E., 1996, MNRAS, 281, 84

van den Bosch F. C., 2000, ApJ, 530, 177

van den Bosch F. C., 2001, MNRAS, 327, 1334

van der Hulst J. M., Skillman E. D., Smith T. R., Bothun G. D., McGaugh S. S., de Blok W. J. G., 1993, AJ, 106, 548

van Kampen E., 1995, MNRAS, 273, 295

van Kampen E., Jimenez R., Peacock J. A., 1999, MNRAS, 310, 43

Walker I. R., Mihos J. C., Hernquist L., 1996, ApJ, 460, 121

Walker T. P., Steigman G., Kang H., Schramm D. M., Olive K. A., 1991, ApJ, 376, 51

Wang B., Heckman T. M., 1996, ApJ, 457, 645

Warren M. S., Quinn P. J., Salmon J. K., Zurek W. H., 1992, ApJ, 399, 405

White S. D. M., Efstathiou G., Frenk C. S., 1993, MNRAS, 262, 1023

White S. D. M., Frenk C. S., 1991, ApJ, 379, 52

White S. D. M., Rees M. J., 1978, MNRAS, 183, 341

Woosley S. E., Weaver T. A., 1995, ApJS, 101, 181

Worthey G., 1994, ApJS, 95, 107

Wyse R. F. G., 2000, in C. M. Carollo, H. C. Ferguson, & R. F. G. Wyse (eds.), *The Formation of Galactic Bulges*, Cambridge: Cambridge University Press, p. 195

Wyse R. F. G., Gilmore G., Franx M., 1997, ARA&A, 35, 637

Young J. S., Knezek P. M., 1989, ApJ, 347, L55

Zhang B., Wyse R. F. G., 2000, MNRAS, 313, 310

Zucca E., et al., 1997, A&A, 326, 477

Zwaan M. A., van der Hulst J. M., de Blok W. J. G., McGaugh S. S., 1995, MNRAS, 273, L35

Appendix A

Plummer softening

Softening is necessary in N-body simulations to prevent excessively strong two-body encounters, which would occur if the particles were treated as point masses with Keplerian potentials, $\phi(r) \propto 1/r$. There are many choices of softened potential to choose from, the exact form of the potential being less important than the overall effect of its presence. The simulations presented in this thesis use Plummer (1911) softening.

The Plummer density profile has the form

$$\rho(r) = \frac{3M}{4\pi} \frac{\epsilon^2}{(r^2 + \epsilon^2)^{5/2}}, \tag{A.1}$$

where ϵ is the softening length. The mass contained within a radius r is

$$M(r) = \frac{M}{(1 + \epsilon^2/r^2)^{3/2}} \tag{A.2}$$

from which we can see that the half-mass radius is

$$r_{\rm half} \simeq 1.3\epsilon. \tag{A.3}$$

The gravitational potential energy of a system described by such a density profile is

$$U = -\frac{3\pi}{32} \frac{GM^2}{\epsilon} \simeq 0.4 \frac{GM^2}{r_{\rm half}}. \tag{A.4}$$

We can use this to re-write the condition for virialisation, $U + 2K = 0$ (where K is the kinetic energy of the system), in terms of the half-mass radius and the velocity dispersion, $\langle v^2 \rangle = 2K/M$:

$$\langle v^2 \rangle \simeq 0.4 \frac{GM}{r_{\rm half}}. \tag{A.5}$$

This justifies the use of equation (2.8) to test groups of particles in the N-body simulation for virialisation.

When a group of particles is replaced by a single softened halo particle with mass

$$M = \sum_i m_i \tag{A.6}$$

and velocity

$$\mathbf{V} = \sum_i \mathbf{v}_i, \tag{A.7}$$

where m_i and $\mathbf{v}_i$ are the masses of the original particles, momentum is conserved but the kinetic energy of the new particle will be less than the sum of the kinetic energies of the original particles:

$$K = \frac{M|\mathbf{V}|^2}{2} < \frac{1}{2}\sum_i m_i|\mathbf{v}_i|^2 \tag{A.8}$$

The remaining energy must go into the internal energy of the new particle. For a virialised system, it is enough to specify the potential energy, since the total energy $E = U/2$. To ensure that energy is (approximately) conserved, the softening length of the new particle must therefore correspond to the half-mass radius of the original group (equation A.3), since we require this to be virialised in the first place.

Appendix B

Halo profiles

B.1 Singular isothermal sphere

The simplest density profile is the singular isothermal sphere, so-called because it has a constant velocity dispersion and is a solution to the equation of hydrostatic support for a self-gravitating sphere of gas. It takes the form

$$\rho(r) = \frac{V_c^2}{4\pi G r^2}, \tag{B.1}$$

where V_c is the circular velocity, defined to be

$$V_c = \left(\frac{GM}{r}\right)^{1/2}. \tag{B.2}$$

Since the circular velocity in this case is constant, the mass diverges at large radii:

$$M(r) = \frac{V_c^2}{G} r \tag{B.3}$$

so to obtain a finite mass the profile needs to be truncated. The binding energy of a virialised isothermal sphere, truncated at a radius $r_{\rm vir}$, is

$$E = \frac{U}{2} = -\frac{GM^2}{2r_{\rm vir}} = -\frac{MV_c^2}{2}, \tag{B.4}$$

where the first equality comes from the virial criterion: $2K + U = 0$.

B.2 'Universal' NFW profile

The haloes identified in N-body simulations are well-fitted by a 'universal' density profile of the form

$$\rho(r) = \frac{\Delta_c \rho_0}{r/r_s(1+r/r_s)^2} \tag{B.5}$$

(Navarro, Frenk & White 1997, NFW), where Δ_c is a characteristic overdensity (ρ_0 here signifies the mean density of the universe) and r_s is a scale radius, at which the density profile changes slope, from $\mathrm{d}\log\rho/\mathrm{d}\log r = -3$ for $r \gg r_s$ to $\mathrm{d}\log\rho/\mathrm{d}\log r = -1$ for $r \ll r_s$. Haloes vary only in the value of this scale radius, r_s. The scale radius is uniquely related to the charateristic overdensity, $\Delta_c(z, z_{\rm coll})$, which in turn is determined by the collapse redshift of the halo. Peacock & Smith (2000) give a method for calculating $\Delta_c(z, z_{\rm coll})$ for a halo of mass M at redshift z (this method improves on that originally employed by NFW).

The mass contained within a radius r ($< r_{\rm vir}$) is

$$M(r) = M\frac{\ln(1+cx) - \frac{cx}{1+cx}}{\ln(1+c) - \frac{c}{1+c}}, \tag{B.6}$$

where $x = r/r_{\rm vir}$ and $c = r_{\rm vir}/r_s$ is the concentration parameter. This leads to a rotation curve of the form

$$V_c(x) = V_c\left[\frac{1}{x}\frac{\ln(1+cx) - \frac{cx}{1+cx}}{\ln(1+c) - \frac{c}{1+c}}\right]^{1/2}. \tag{B.7}$$

The binding energy of a halo with a NFW profile can be calculated from the rotation curve:

$$\begin{aligned} E = \frac{U}{2} &= -2\pi r_{\rm vir}^3 \int_0^1 x^2 \rho(x) V_c^2(x)\,\mathrm{d}x \\ &= -\frac{GM^2}{2r_{\rm vir}} f_c. \end{aligned} \tag{B.8}$$

Mo, Mao & White (1998) give an approximate expression (accurate to within 1% for $5 < c < 30$) for f_c:

$$f_c \simeq \frac{2}{3} + \left(\frac{c}{21.5}\right)^{0.7}. \tag{B.9}$$

B.3 Pseudo-isothermal (core) profile

It is commonly assumed that, in the absence of cooling, galaxy haloes are well-described by equation (B.5), as they are in N-body simulations. Observationally, however, there is little evidence for dark matter profiles with strong central cusps, a feature common to both the NFW profile and the singular isothermal sphere. The rotation curves of low surface-brightness dwarf galaxies are better fit by profiles with constant density cores (de Blok & McGaugh 1997; Swaters, Madore & Trewhella 2000; de Blok & Bosma 2002). One such profile is the pseudo-isothermal, or core, profile:

$$\rho(r) = \frac{\rho_0}{1 + (r/r_\mathrm{c})^2}, \tag{B.10}$$

where ρ_0 is the (constant) central density and r_c is the core radius. The mass contained within radius r is

$$M(r) = Mx\frac{1 - \tan^{-1}(ax)/ax}{1 - \tan^{-1}(a)/a}, \tag{B.11}$$

where, in analogy with the NFW c parameter, I define a core parameter: $a = r_\mathrm{vir}/r_\mathrm{c}$. The core profile has a rotation curve of the form

$$V_\mathrm{c}(x) = V_\mathrm{c}\left[\frac{1 - \tan^{-1}(ax)/ax}{1 - \tan^{-1}(a)/a}\right]^{1/2}. \tag{B.12}$$

The binding energy,

$$E = -2\pi r_\mathrm{vir}^3 \int_0^1 x^2 \rho(x) V_\mathrm{c}^2(x)\,\mathrm{d}x, \tag{B.13}$$

has to be integrated numerically in this case. This procedure can be time-consuming, so it is useful to search for an approximation similar to equation B.9. Following the approach of Mo, Mao & White, I define a factor f_a, such that

$$E = -\frac{GM^2}{2r_\mathrm{vir}} f_\mathrm{a}. \tag{B.14}$$

Performing the integration numerically, I find that the approximation

$$f_\mathrm{a} \simeq 0.5 + \left(\frac{a}{3300}\right)^{0.2} \tag{B.15}$$

is sufficiently accurate, to within 1% for $10 \leqslant a \leqslant 40$.

There is no theoretical prediction for the core sizes of galaxy haloes; these have to be determined empirically, from observations. The core radii of galaxy-scale haloes

– inferred from fits to the rotation curves of dwarf galaxies – and those of clusters – measured using gravitational lensing for a single cluster – are consistent with a relation of the form $r_c \propto V_c$ (Firmani et al. 2001), implying that $r_c \propto r_{vir}$, so that a single value of a is appropriate for haloes of all masses.

Appendix C

Measurements in cosmology

C.1 Angular diameter distance

The proper tranverse size, $\mathrm{d}l$, of an object of angular size $\mathrm{d}\psi$ on the sky is given by

$$\begin{aligned} \mathrm{d}l &= RS_k(r)\,\mathrm{d}\psi \\ &= (1+z)^{-1}R_0S_k(r)\,\mathrm{d}\psi. \end{aligned} \tag{C.1}$$

This leads us to define an *angular diameter distance*, D_A, by requiring that

$$\mathrm{d}l = D_\mathrm{A}\mathrm{d}\psi, \tag{C.2}$$

as in the classical case. This leads to

$$D_\mathrm{A} = (1+z)^{-1}R_0S_k(r) \tag{C.3}$$

The line-of-sight comoving distance, r, to the object can be found by integrating the metric along a radial null trajectory:

$$r = \int_0^r \mathrm{d}r' = c\int_0^t \frac{\mathrm{d}t'}{R(t')}. \tag{C.4}$$

Using the definition of the Hubble parameter $H(R)$, this can be written in terms of redshifts as

$$r = \frac{c}{R_0}\int_0^z \frac{\mathrm{d}z'}{H(z')}, \tag{C.5}$$

where $H(z)$ is given by equation (1.13).

C.2 Luminosity distance

The broadband flux, $S_{\rm BB}$, of an object, integrated over some range of wavelengths measured in the observers frame, can be written as

$$S_{\rm BB} = \frac{L_{\rm BB}}{4\pi R_0^2 S_k^2(r)(1+z)^2}, \tag{C.6}$$

where $L_{\rm BB}$ is the luminosity of the object integrated over the corresponding range of rest-frame wavelengths. As above, we can define a *luminosity distance* such that the classical result,

$$S_{\rm BB} = \frac{L_{\rm BB}}{4\pi D_{\rm L}^2}, \tag{C.7}$$

holds. So

$$\begin{aligned} D_{\rm L} &= (1+z)R_0 S_k(r) \\ &= (1+z)^2 D_{\rm A}. \end{aligned} \tag{C.8}$$

C.3 Surface brightness

The integrated surface brightness, I, of a uniformly illuminated object of solid angle $\mathrm{d}\Omega$ is

$$\begin{aligned} I &= \frac{\mathrm{d}S_{\rm BB}}{\mathrm{d}\Omega} \\ &= \frac{\mathrm{d}L_{\rm BB}/4\pi D_{\rm L}^2}{\mathrm{d}A/D_{\rm A}^2}, \end{aligned} \tag{C.9}$$

where $\mathrm{d}A$ is the proper area of the source. The intrinsic surface brightness of the source is just

$$I_0 = \frac{\mathrm{d}L_{\rm BB}}{4\pi\,\mathrm{d}A} \tag{C.10}$$

so equation C.9 simplifies to

$$I = I_0(1+z)^{-2}. \tag{C.11}$$

So, at low redshifts ($z \ll 1$) surface brightness is conserved, as in the classical case, while at cosmologically significant distance, there is a factor $(D_{\rm A}/D_{\rm L})^2$ dimming.

C.4 Comoving volume

The comoving volume element at comoving distance $R_0 r$ corresponding to a solid angle $\mathrm{d}\Omega$ is

$$\begin{aligned} \mathrm{d}V &= R_0\,\mathrm{d}r \times [R_0 S_k(r)]^2\,\mathrm{d}\Omega \\ &= R_0\,\mathrm{d}r \times (1+z)^2 D_\mathrm{A}^2\,\mathrm{d}\Omega, \end{aligned} \tag{C.12}$$

where D_A is the angular diameter distance, given by equation (C.3). Using equation (C.5), this can be written in terms of the redshift element, $\mathrm{d}z$:

$$\mathrm{d}V = \frac{c(1+z)^2 D_\mathrm{A}^2}{H(z)}\,\mathrm{d}\Omega\,\mathrm{d}z. \tag{C.13}$$

Appendix D

Galaxy photometry

D.1 Discs

The surface brightness profile of an exponential disc is characterised by the central surface brightness, μ_0, and an exponential scale length, h, in angular units. In magnitude units, the exponential profile takes the simple form

$$\mu(R) = \mu_0 + 1.086(R/h), \tag{D.1}$$

where the face-on central surface brightness is related to the total magnitude, m, by

$$\mu_0 = m + 2.5\log_{10}(2\pi h^2). \tag{D.2}$$

The fraction of the flux contained within a radius R is

$$f(R) = 1 - (1 + R/h)\mathrm{e}^{-R/h}. \tag{D.3}$$

From this, the half-light radius can be found numerically, using an iterative procedure, leading to $R_\mathrm{e} = 1.678h$. The mean surface-brightness within this radius is thus

$$\mu_\mathrm{e} = m + 2.5\log_{10}(2\pi R_\mathrm{e}^2) = \mu_0 + 1.124. \tag{D.4}$$

This is also commonly known as the *effective* surface-brightness (see below). For a pure disc galaxy, the isophotal radius for a limiting isophote μ_lim is given by

$$R_\mathrm{iso} = 0.921h(\mu_\mathrm{lim} - \mu_0) \tag{D.5}$$

and the isophotal magnitude by

$$m_{\rm iso} = m_{\rm tot} - 2.5 \log_{10} f(R_{\rm iso}), \tag{D.6}$$

where $f(R)$ is given by equation (D.3).

For a disc with inclination i to the line of sight, the element of area is reduced by a factor $\cos i$, so the surface brightness is increased by the inverse of this factor.

D.2 Bulges

I assume an $R^{1/4}$–profile:

$$\mu(R) = \mu_{\rm e} + 8.33[(R/R_{\rm e})^{1/4} - 1] + 1.39 \tag{D.7}$$

(de Vaucouleurs' 1948) for spheroids of all sizes. Here, $R_{\rm e}$ is the half-light, or *effective*, radius and $\mu_{\rm e}$ is the mean surface brightness within this radius. Some care needs to be taken here, because in another commonly used notation, the effective surface brightness, $\mu_{\rm e}$, is the surface brightness *at* $R_{\rm e}$:

$$\mu(R_{\rm e}) = \mu_{\rm e} + 1.39. \tag{D.8}$$

The central surface brightness of the $R^{1/4}$–profile is clearly

$$\mu_0 = \mu_{\rm e} - 6.94. \tag{D.9}$$

The effective surface brightness is related to the apparent magnitude, m, by

$$\mu_{\rm e} = m + 2.5 \log_{10}(2\pi R_{\rm e}^2). \tag{D.10}$$

The fraction of the flux contained within a radius R in this case is given, approximately, by

$$g(R) = 1 - \sum_{i=0}^{7} x^i {\rm e}^x \quad with \quad x = 7.670(R/R_{\rm e})^{1/4}. \tag{D.11}$$

For an elliptical galaxy with an $R^{1/4}$–profile, the isophotal radius is

$$R_{\rm iso} = R_{\rm e}[0.12(\mu_{\rm lim} - \mu_{\rm e}) + 1]^4 \tag{D.12}$$

and the isophotal magnitude is

$$m_{\rm iso} = m_{\rm tot} - 2.5 \log_{10} g(R_{\rm iso}). \tag{D.13}$$